

Continuous Mathematical Methods, emphasizing Machine Learning

Ron Fedkiw^{*1}, Yilin Zhu^{*23}, Winnie Lin^{*3}, Jane Wu^{*3}

^{*} Stanford University, ¹ Slide Design and Content (and Instructor), ² Slide Illustrator, ³ Teaching Assistant

Table of Contents

- Unit 1: Introduction
- Unit 2: Linear Systems
- Unit 3: Understanding Matrices
- Unit 4: Special Matrices
- Unit 5: Iterative Solvers
- Unit 6: Local Approximations
- Unit 7: Curse of Dimensionality
- Unit 8: Least Squares
- Unit 9: Basic Optimization
- Unit 10: Solving Least Squares
- Unit 11: Zero Singular Values
- Unit 12: Regularization
- Unit 13: Optimization
- Unit 14: Nonlinear Systems
- Unit 15: 1D Root Finding
- Unit 16: 1D Optimization
- Unit 17: Computing Derivatives
- Unit 18: Avoiding Derivatives
- Unit 19: Descent Methods
- Unit 20: Momentum Methods
- Appendix: Notation

Unit 1

What is Learning?

What is Learning?

- There are lots of answers to this question, and explanations often become philosophical
- A more practical question might be:

What can we teach/train a person, animal, or machine to do?

Example: Addition “+”

- How is addition taught in schools?
 - Memorize rules for pairs of numbers from the set $\{0,1,2,3,4,5,6,7,8,9\}$
 - Memorize redundant rules for efficiency, e.g. $0+x=x$
 - Learn to treat powers of 10 implicitly, e.g. $12+34=46$ because $1+3=4$ and $2+4=6$
 - Learn to carry when the sum of two numbers is larger than 9
 - Learn to add larger sets of numbers by considering them one pair at a time
 - Learn how to treat negative numbers
 - Learn how to treat decimals and fractions

Knowledge Based Systems (KBS)

Contains two parts:

1) Knowledge Base

- Explicit knowledge or facts
- Often populated by an expert (expert systems)

2) Inference Engine

- Way of reasoning about the facts in order to generate new facts
- Typically follows the rules of Mathematical Logic

See Wikipedia for more details...

KBS Approach to Addition

- Rule: x and y commute
- Start with x and y as single digits, and record all $x + y$ outcomes as facts (using addition)
- Add rules to deal with numbers with more than one digit by pulling out powers of 10
- Add rules for negative numbers, decimals, fractions, etc.
- Mimics human learning (or at least human **teaching**)
- This is a discrete approach, and it has no inherent error!

Machine Learning (ML)

Contains two parts:

1) Training Data

- Data Points - typically as domain/range pairs
- Hand labeled by a user, measured from the environment, or generated procedurally

2) Model

- Derived from Training Data in order to estimate new data points minimizing errors
- Uses Algorithms, Statistical Reasoning, Rules, Networks, Etc.

See Wikipedia for more details...

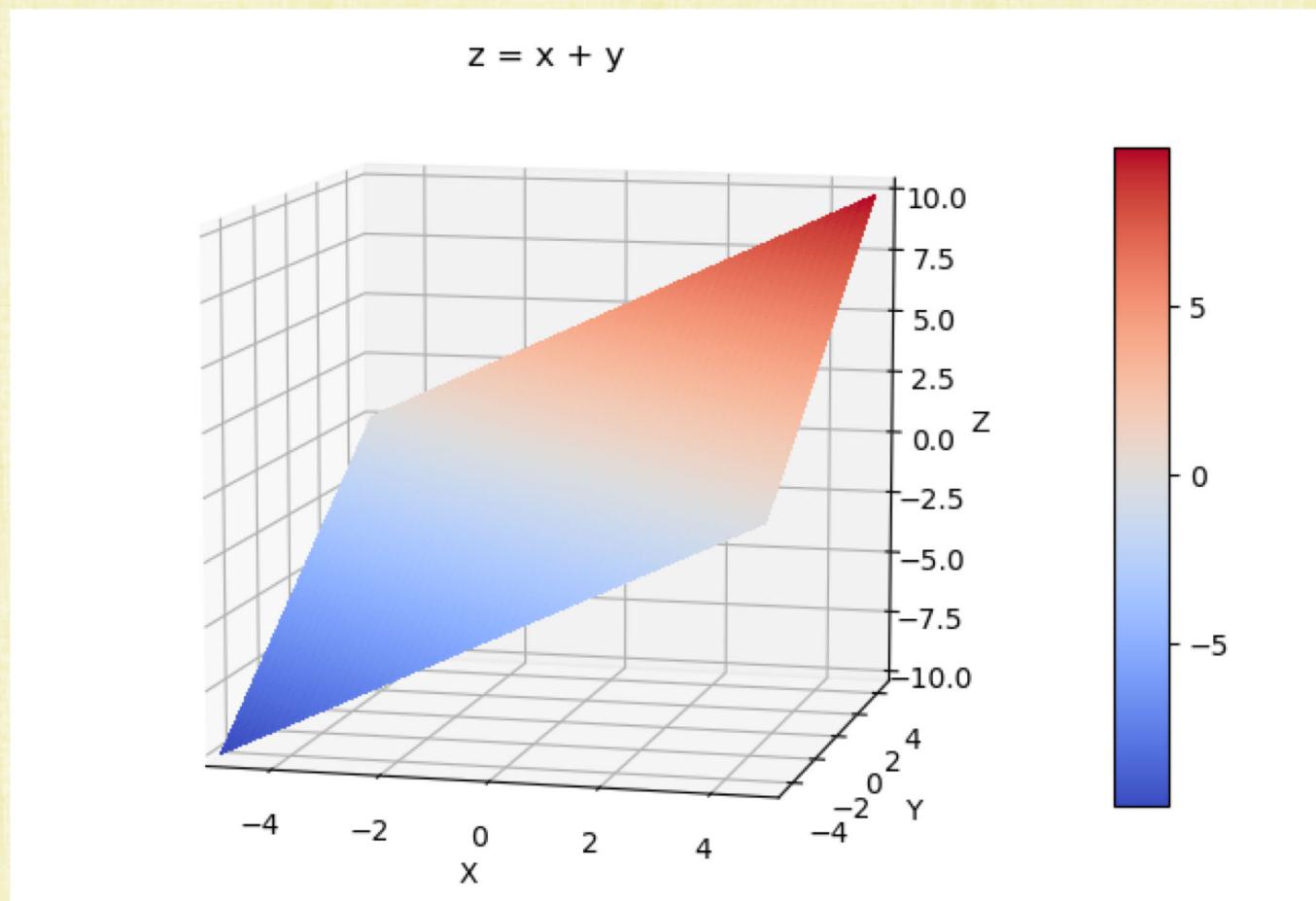
KBS vs. ML

- KBS and ML can be seen as the discrete math and continuous math approaches (respectively) to the same problem
- ML's Training Data serves the same role as KBS's Knowledge Base
- Logic is the algorithm used to discover new discrete facts for KBS, whereas many algorithms/methods are used to approximate continuous facts/data for ML
 - Logic (in particular) happens to be especially useful for discrete facts
- ML, derived from continuous math, will tend to have inherent approximation errors

ML Approach to Addition

- Make a $2D$ domain in R^2 , and a $1D$ range R^1 for the addition function
- As training data, choose a number of input points (x_i, y_i) with output $x_i + y_i$
- Plot the $3D$ points $(x_i, y_i, x_i + y_i)$ and determine a model function $z = f(x, y)$ that best approximates the data
- Turns out that the plane $z = x + y$ exactly fits the data
 - only need 3 training points to determine this plane
- Don't need special rules for negative numbers, fractions, irrationals such as $\sqrt{2}$ and π , etc.
- However, small errors in the inputs lead to a slightly incorrect plane, which can cause quite large errors far away from the input data
- This can be alleviated to some degree by using a lot of points and the plane that best fits those points

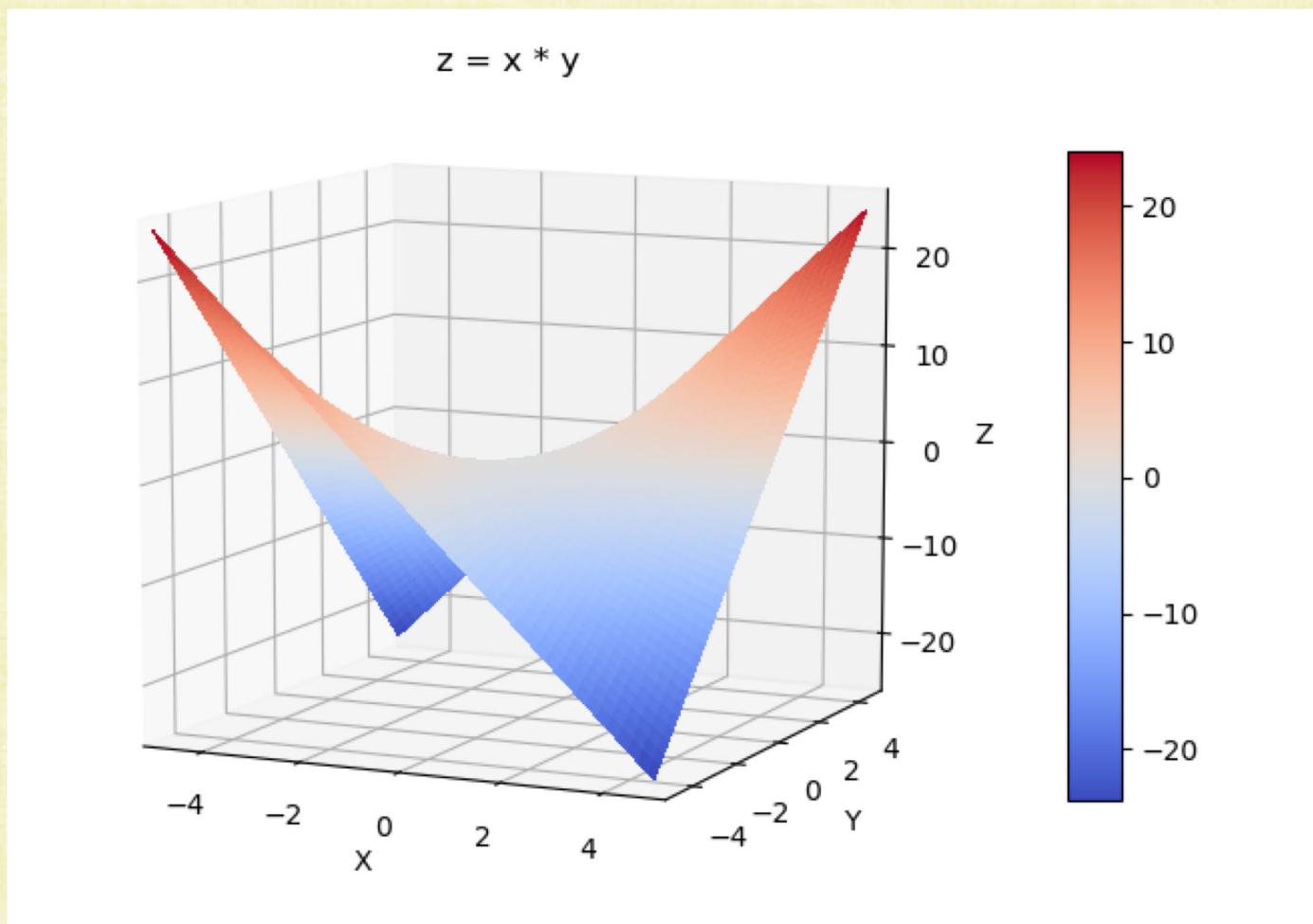
ML Approach to Addition



Example: Multiplication “*”

- KBS creates new rules for $x * y$, utilizing the rules from addition too
- ML utilizes a set of 3D points $(x_i, y_i, x_i * y_i)$ as training data, and the model function $z = x * y$ can be found to exactly fit the data
- However, we are “cheating” by using an inherently represented floating point operation as our model

ML Approach to Multiplication



Example: Unknown Operation “#”

- KBS fails!
- How can KBS create rules for $x\#y$ when we don't know what # means?
- This is the case for many real-world phenomena that are not fully understood
- However, sometimes it is possible to get some examples of $x\#y$
- That is, through experimentation or expert knowledge, can discover $z_i = x_i \# y_i$ for some number of pairs (x_i, y_i)
- Subsequently, these known (or estimated) 3D points (x_i, y_i, z_i) can be used as training data to determine a model function $z = f(x, y)$ that approximately fits the data

Determining the Model Function

- How does one determine $z = f(x, y)$ near the training data, so that it robustly predicts/infers new \hat{z} from inputs (\hat{x}, \hat{y}) not contained in the training data?
- How does one minimize the effect of inaccuracies or noise in the training data?
- Caution: away from the training data, the model function $z = f(x, y)$ is likely to be highly inaccurate (**extrapolation is ill-posed**)

Nearest Neighbor

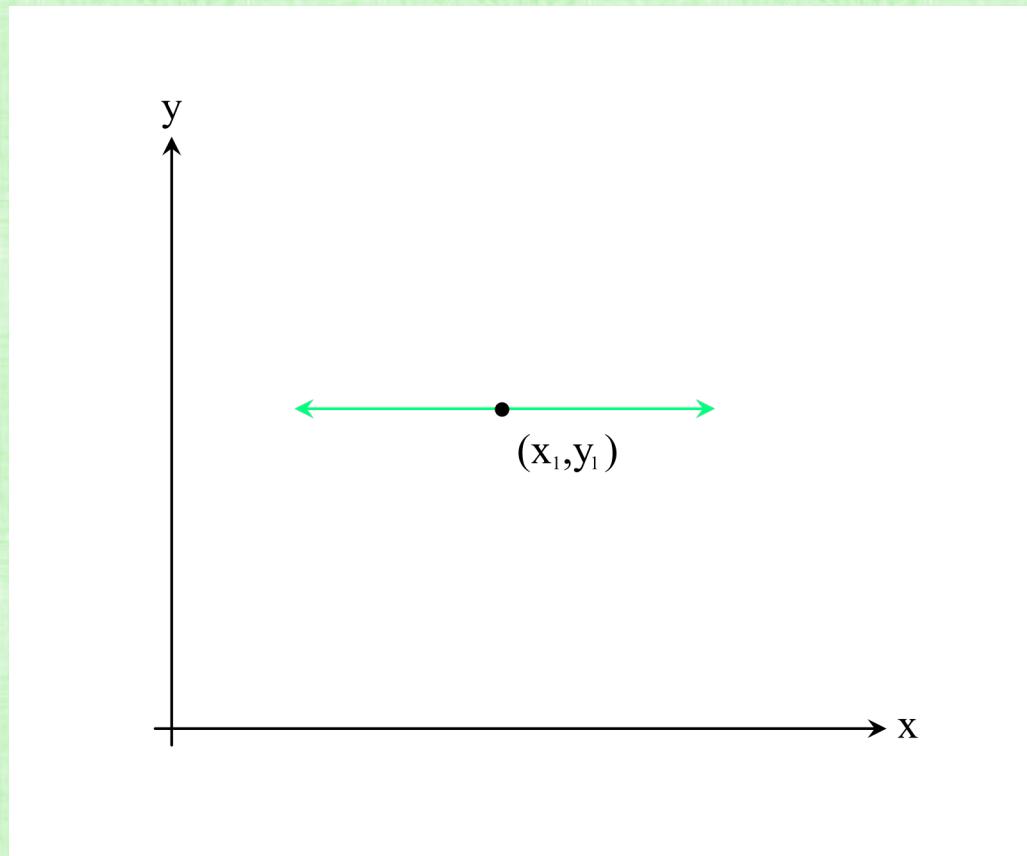
- If asked to multiply 51.023 times 298.5, one might quickly estimate that 50 times 300 is 15,000
- That is a nearest neighbor style algorithm, relying on nearby data where the answer is known, better known, or more easy to come by
- Given (\hat{x}, \hat{y}) , find the closest (Euclidean distance) training data (x_i, y_i) and return the associated z_i (with error $\|z_i - \hat{z}\|$)
- This represents $z = f(x, y)$ as a piecewise constant function with discontinuities on the boundaries of the associated Voronoi regions
- This is the simplest possible Machine Learning algorithm (a piecewise constant function), and it works in an arbitrary number of dimensions

Data Interpolation

- In order to better elucidate various concepts, we consider the interpolation of data in more detail
- We begin by reverting back to the simplest possible case with $1D$ inputs and $1D$ outputs, i.e. $y = f(x)$

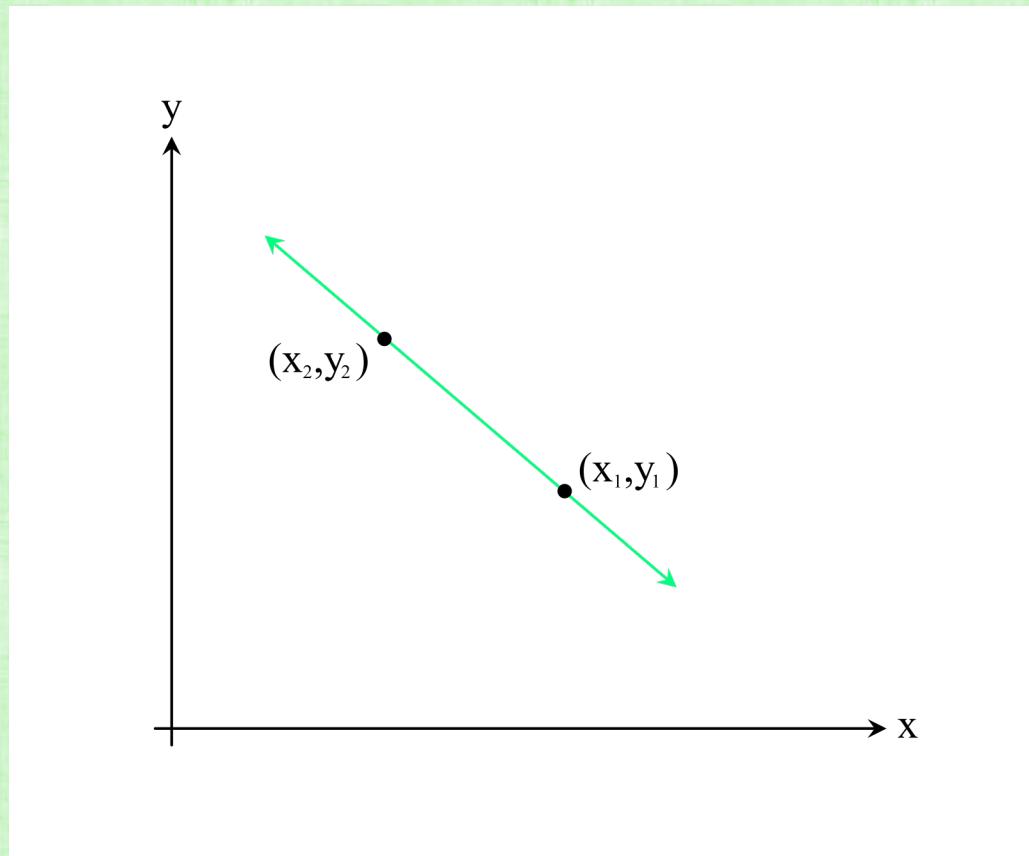
Polynomial Interpolation

- Given 1 data point, one can at best draw a constant function



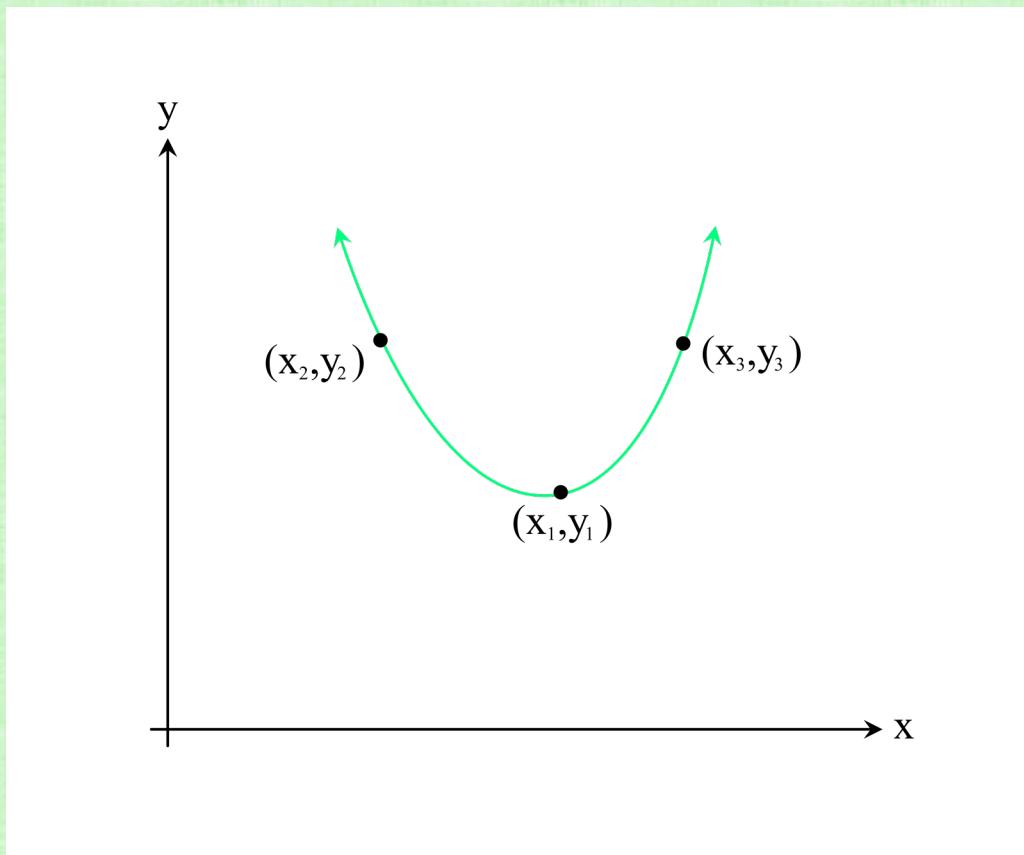
Polynomial Interpolation

- Given 2 data points, one can at best draw a linear function



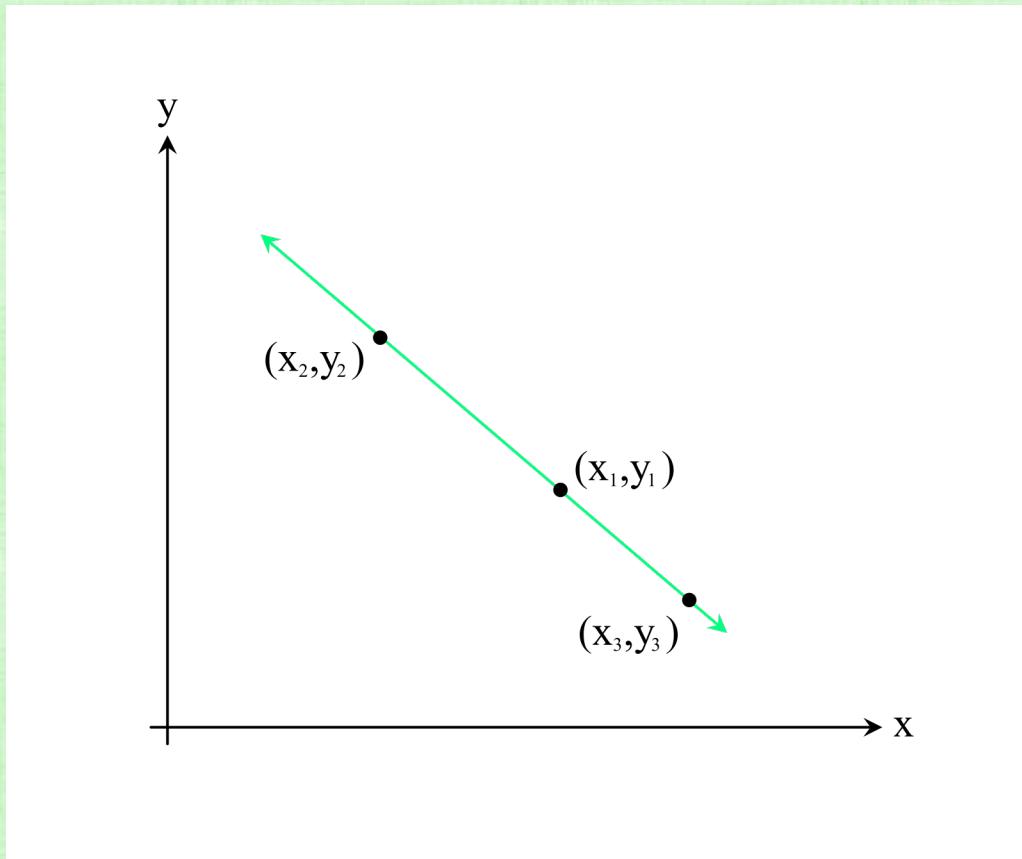
Polynomial Interpolation

- Given 3 data points, one can at best draw a quadratic function



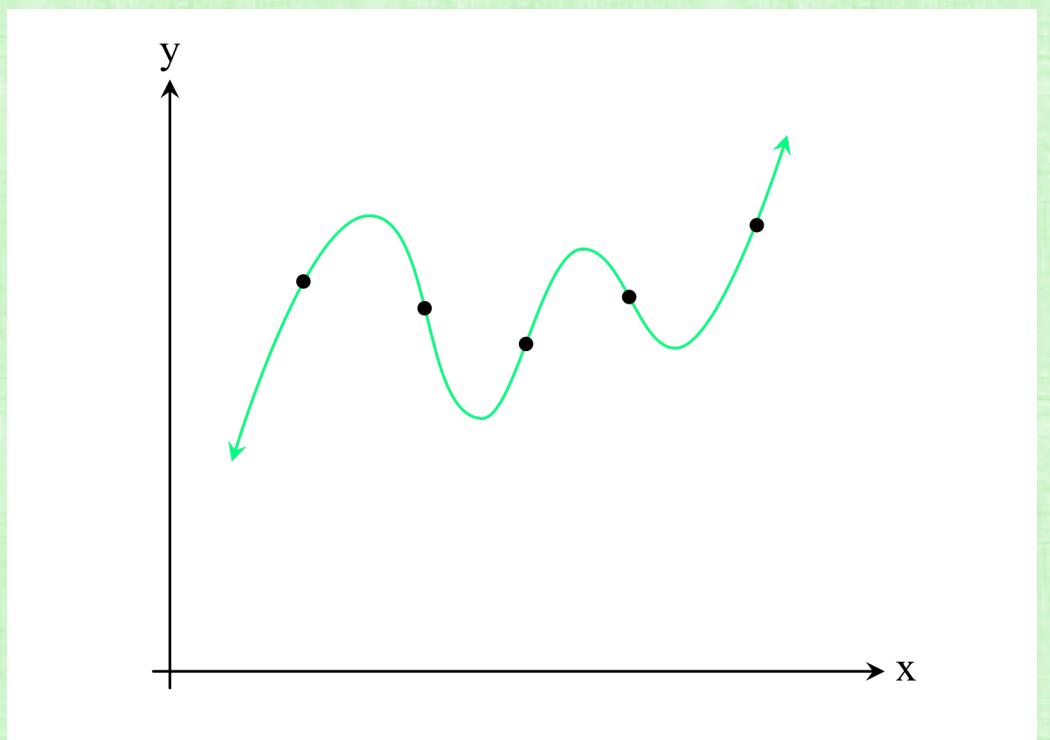
Polynomial Interpolation

- Unless all 3 points are on the same line, in which case one can only draw a linear function



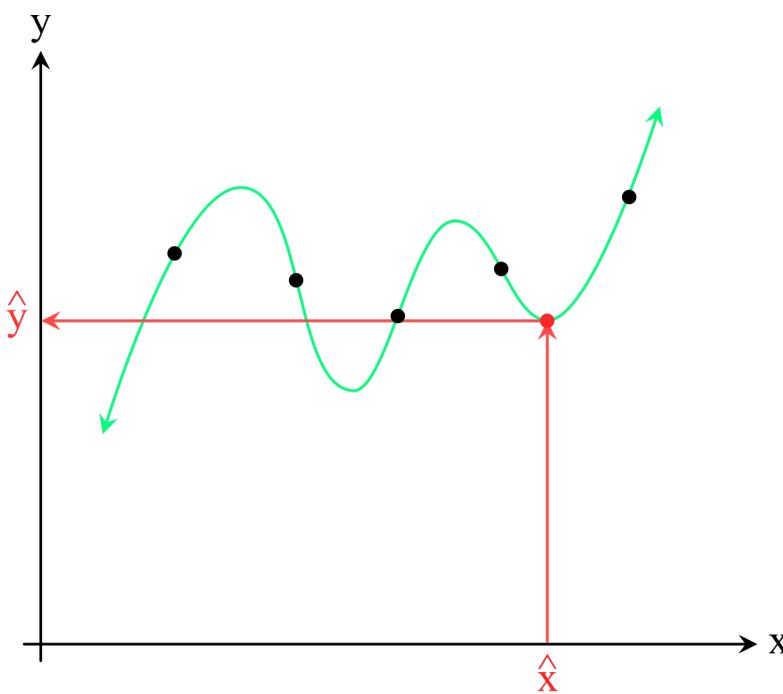
Polynomial Interpolation

- Given m data points, one can draw a unique $m - 1$ degree polynomial that goes through all of them
 - As long as they are not degenerate, like 3 points on a line



Overfitting

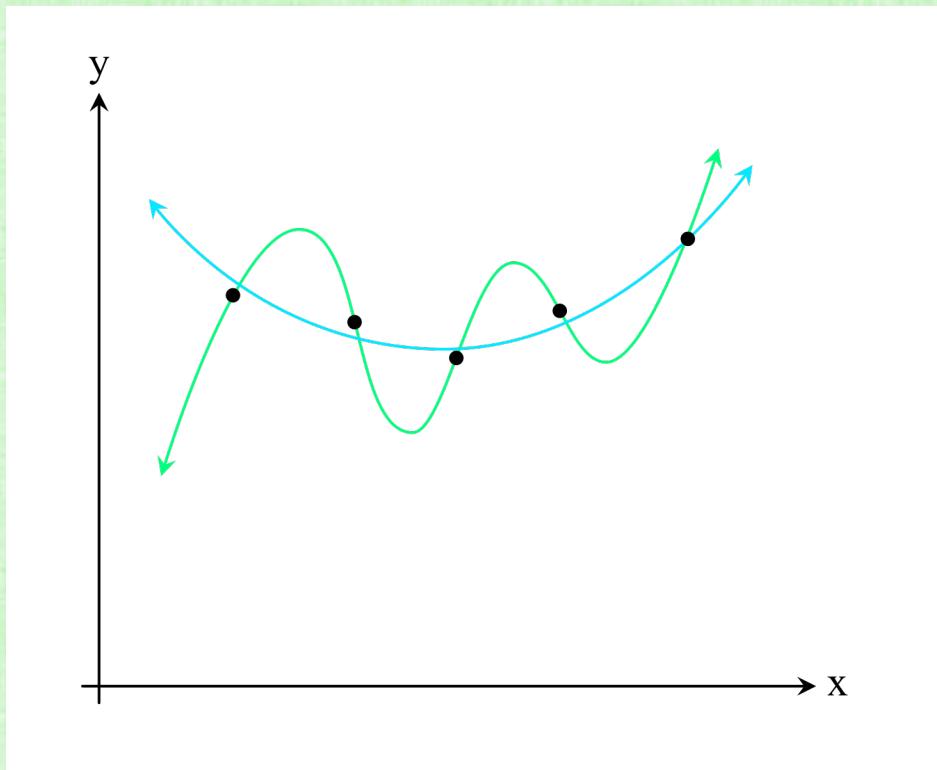
- Given a new input \hat{x} , this interpolant infers/predicts an output \hat{y} that may be far from what one may expect



- Interpolating polynomials are smooth (continuous function and derivatives)
- Thus, they wiggle/overshoot in between data points (so that they can smoothly turn back and hit the next point)
- Overly forcing polynomials to exactly hit every data point is called overfitting (overly fitting to the data)
- It results in inference/predictions that vary too wildly from the training data

Regularization

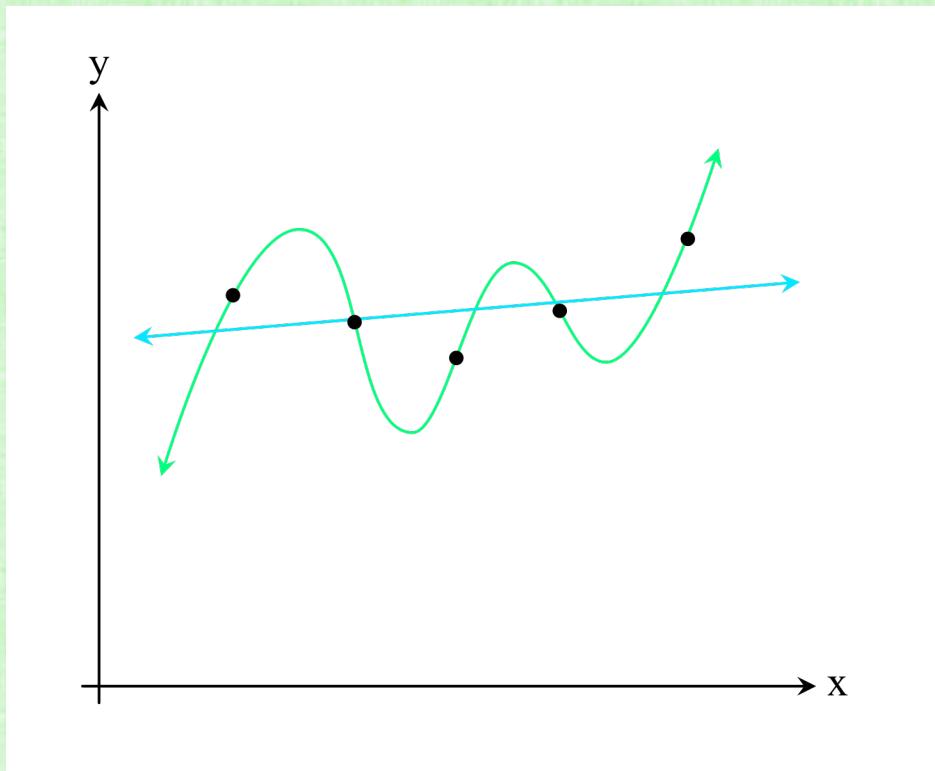
- Using a lower order polynomial that doesn't (can't) exactly fit the data points provides some degree of regularization



- A regularized interpolant contains intentional errors in the interpolant missing some/all the data points
- However, this hopefully makes the function more predictable/smooth between data points
- Moreover, the data points themselves may contain noise/error, so it is not clear whether they should be interpolated exactly

Underfitting

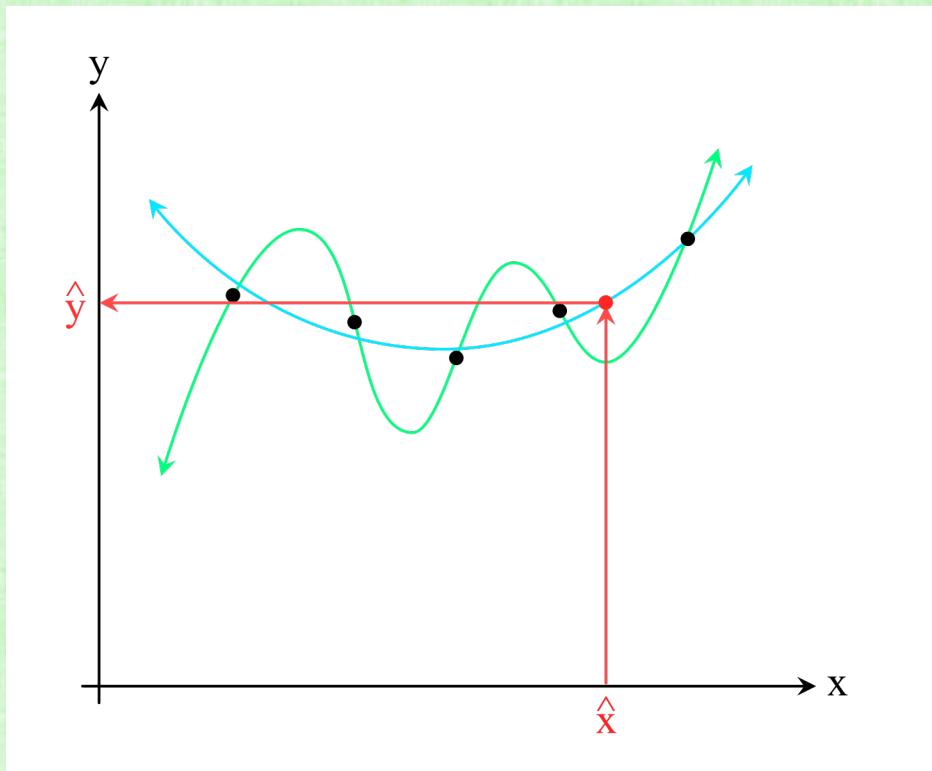
- Using too low of an order polynomial causes one to miss the data by too much



- A linear function doesn't capture the essence of this data as well as a quadratic function does
- Choosing too simple of a model function or regularizing too much prevents one from properly interpolating the data

Regularization

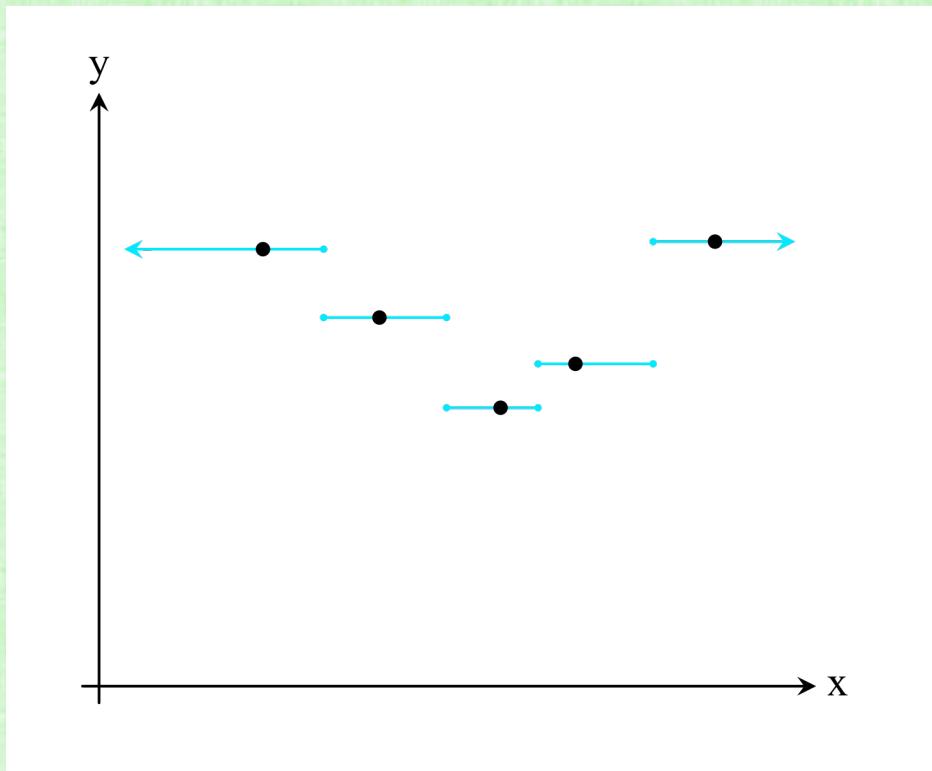
- Given \hat{x} , the regularized interpolant infers/predicts a more reasonable \hat{y}



- There is a trade-off between sacrificing accuracy on fitting the input data, and obtaining better accuracy on inference/prediction for new inputs

Nearest Neighbor

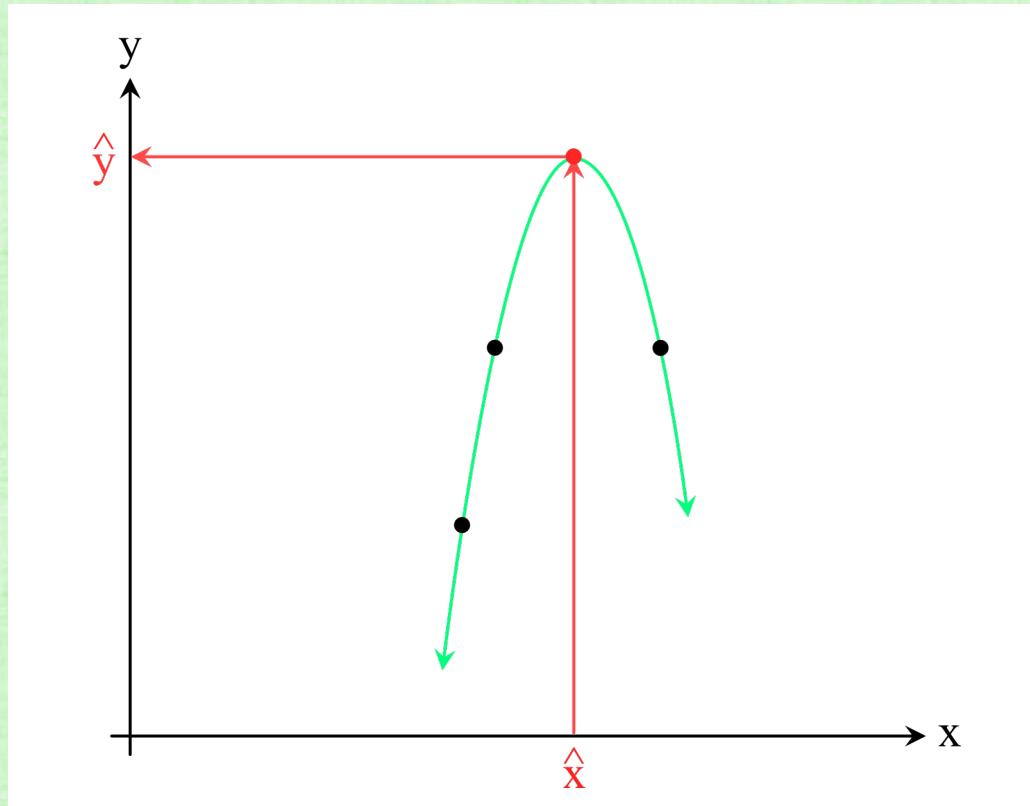
- Piecewise-constant interpolation on this data (equivalent to nearest neighbor)



- The good behavior of the piecewise constant function stresses the importance of approximating data locally
- We will address Local Approximations later in the quarter

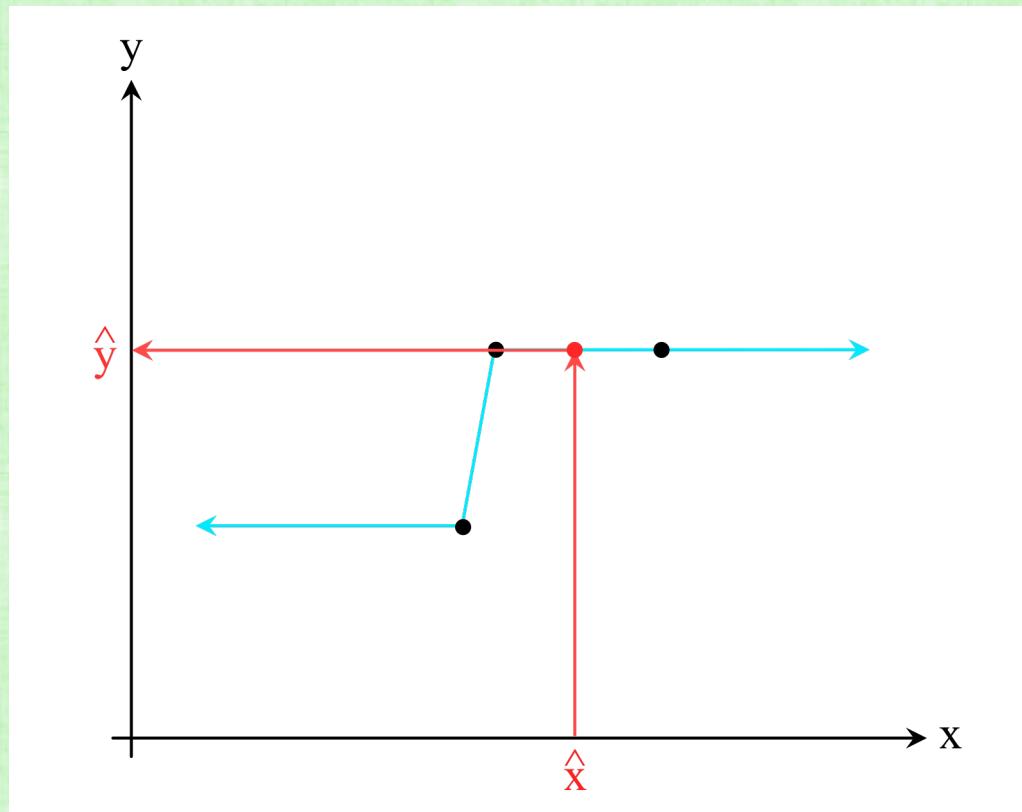
Overfitting

- Although higher order polynomials tend to oscillate more wildly, even a quadratic polynomial can overfit quite a bit



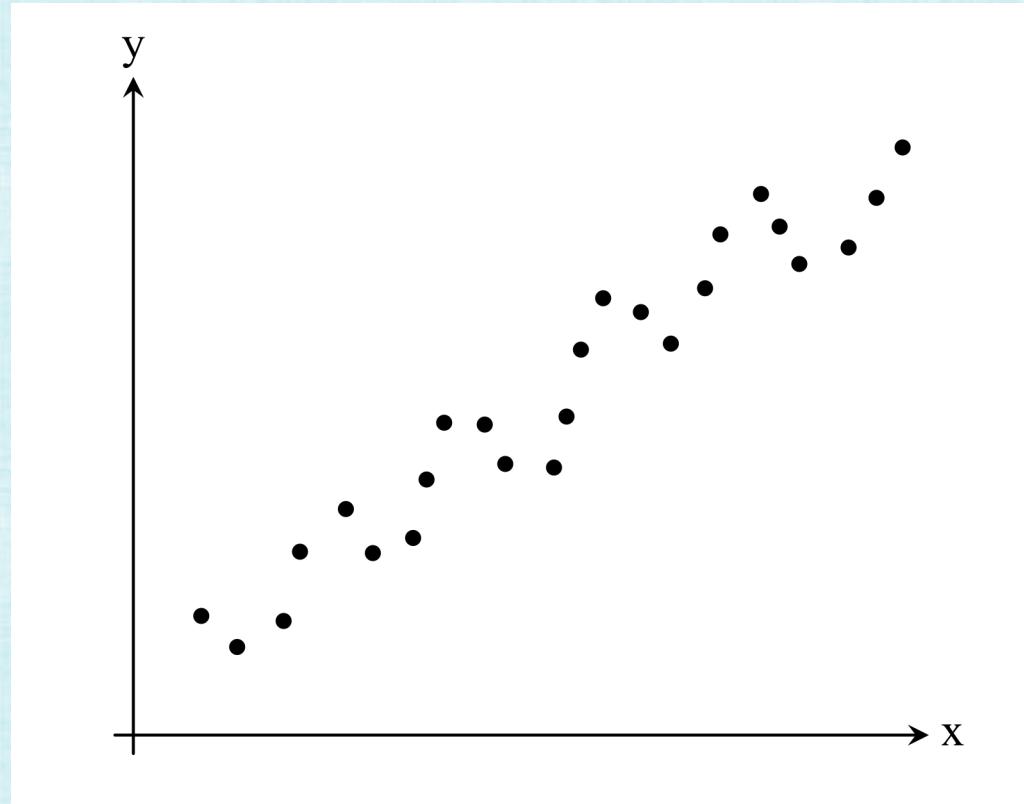
Overfitting

- A piecewise linear approach works better on this data



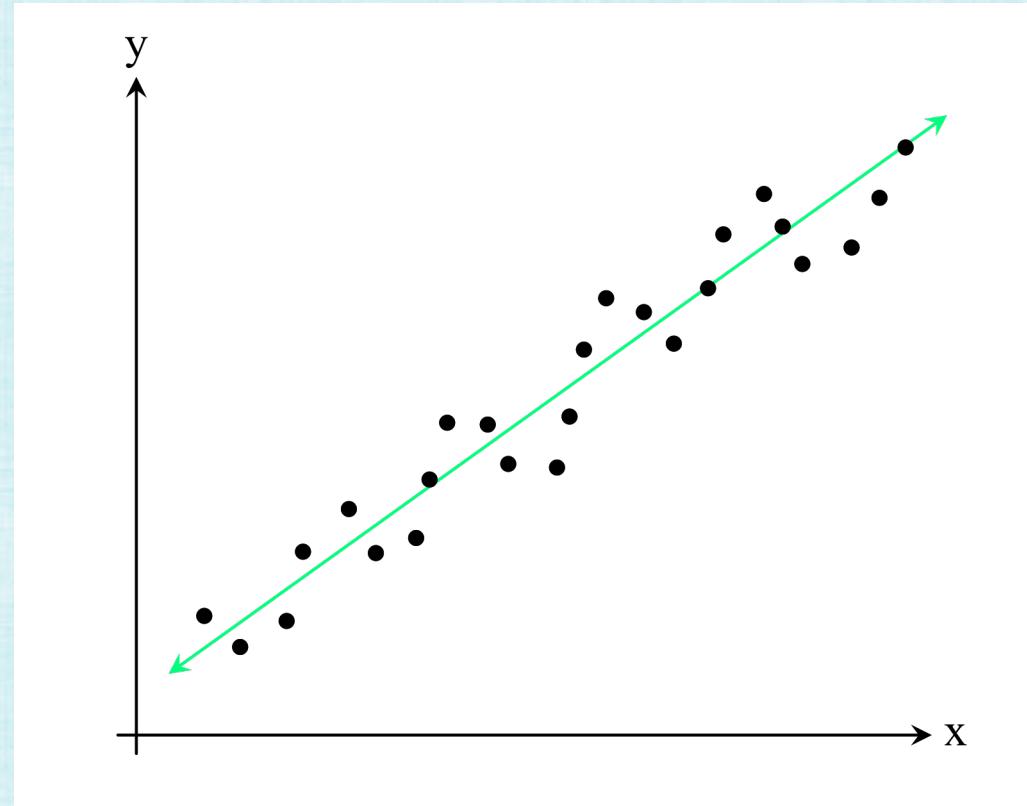
Noisy Data

- There may be many sources of error in data, so it can often be unwise to attempt to fit data too closely



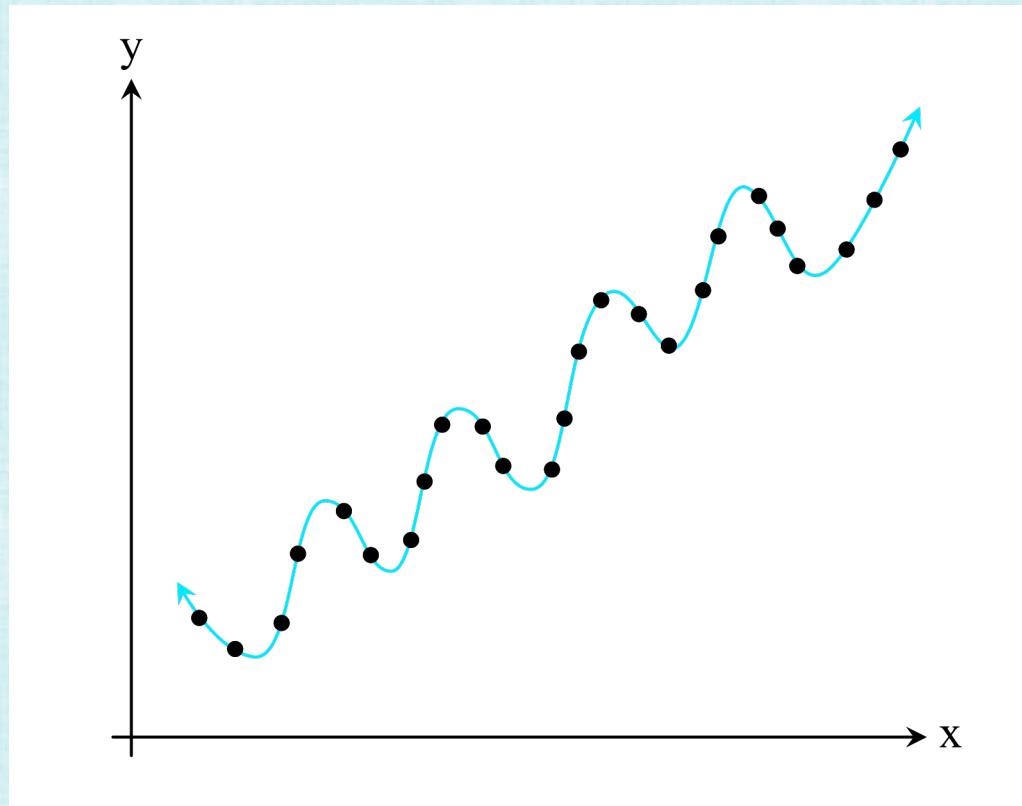
Linear Regression

- One commonly fits a low order model to such data, while minimizing some metric of mis-interpolating data



Noise vs. Features

- But how can one differentiate between noise and features?



Noise vs. Features

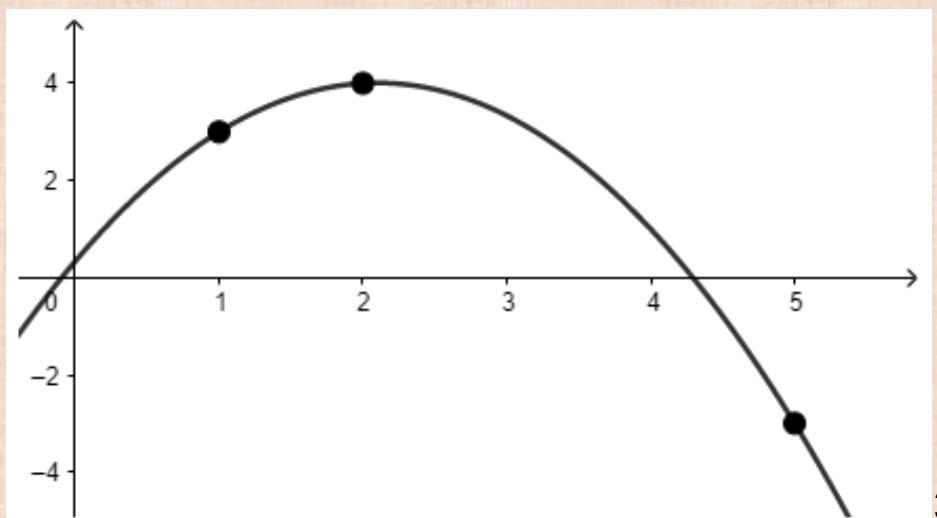
- When training a neural network, split the available data into 3 sets
- E.g., 80% training data, 10% model validation data, and 10% test data
- Training data is used to train the neural network
 - An interpolatory function is fit to that data (potentially overfitting it)
- When considering features vs. noise, overfitting, etc., model validation data is used to select the best model function version or fitting strategy
- Finally, when disseminating results advocating the “best” model, inferencing on the test data gives some idea as to how well that model might generalize to unseen data

Monomial Basis for Polynomial Interpolation

- Given m data points (x_i, y_i) , find the unique polynomial that passes through them: $y = c_1 + c_2x + c_3x^2 + \cdots + c_m x^{m-1}$
- Write an equation for each data point, note that the equations are linear, and put into matrix form
- For example, consider $(1,3), (2,4), (5, -3)$ and a quadratic polynomial

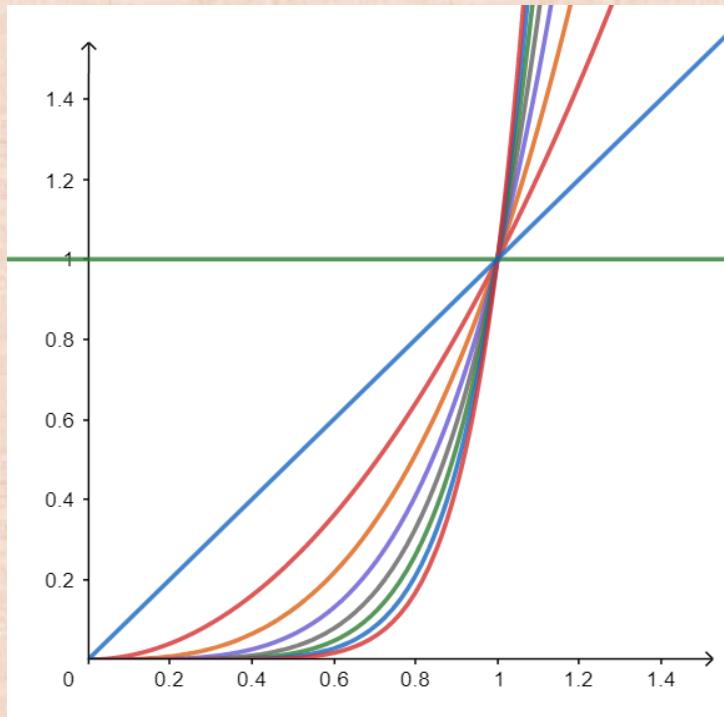
Here, $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 5 & 25 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ -3 \end{pmatrix}$ gives

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 1/3 \\ 7/2 \\ -5/6 \end{pmatrix} \text{ and } f(x) = \frac{1}{3} + \frac{7}{2}x - \frac{5}{6}x^2$$



Monomial Basis for Polynomial Interpolation

- In general, solve $Ac = y$ where A (the Vandermonde matrix) has a row for each data point of the form $(1 \quad x_i \quad x_i^2 \quad \dots \quad x_i^{m-1})$

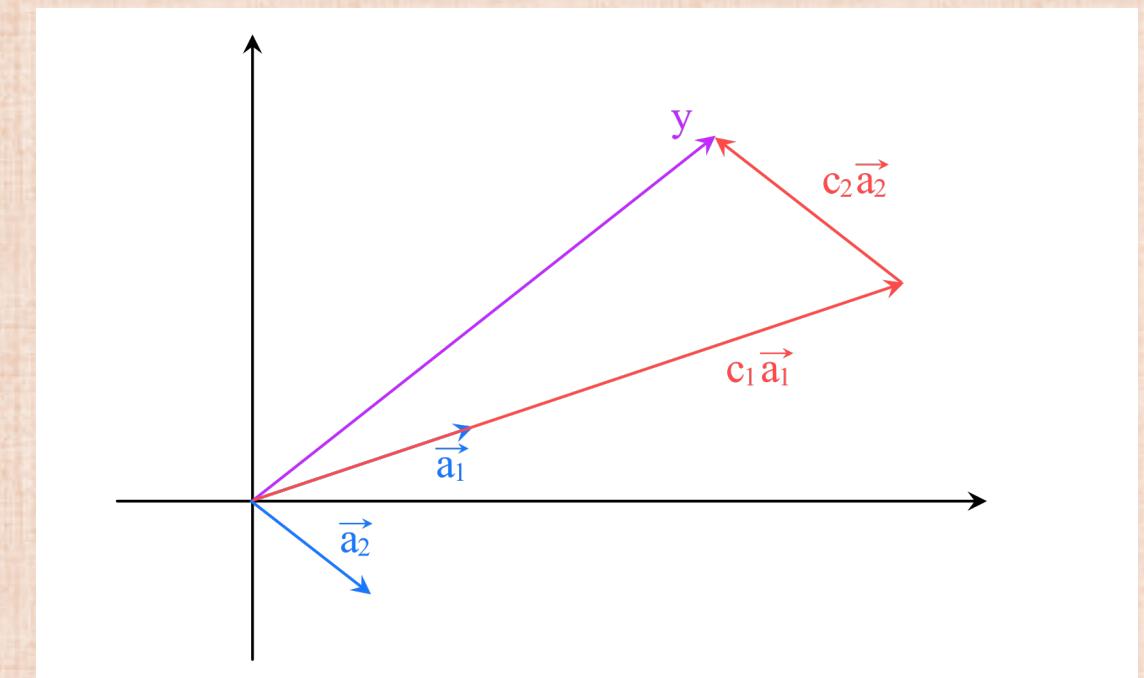
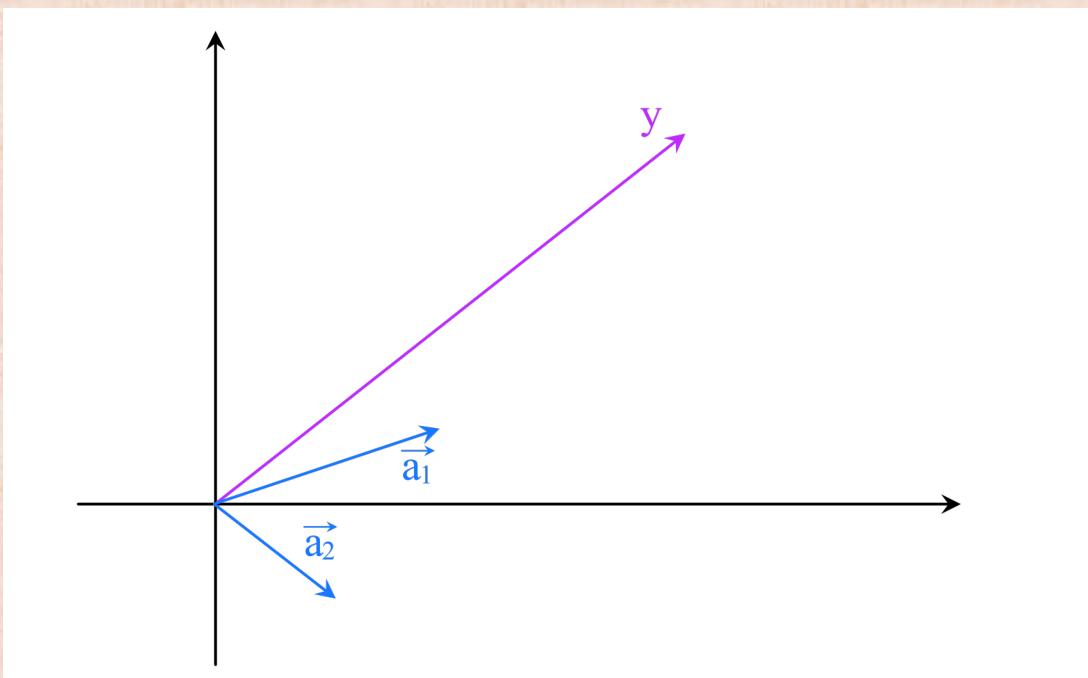


- Polynomials look more similar at higher powers
- This makes the rightmost columns of the Vandermonde matrix tend to become more parallel
- Round-off errors and other numerical approximations exacerbate this
- More parallel columns make the matrix less invertible, and harder to solve for the parameters c_k
- Too nearly parallel columns make the matrix ill-posed and unsolvable on the computer

$$f(x) = 1, x, x^2, x^3, x^4, x^5, x^6, x^7, x^8$$

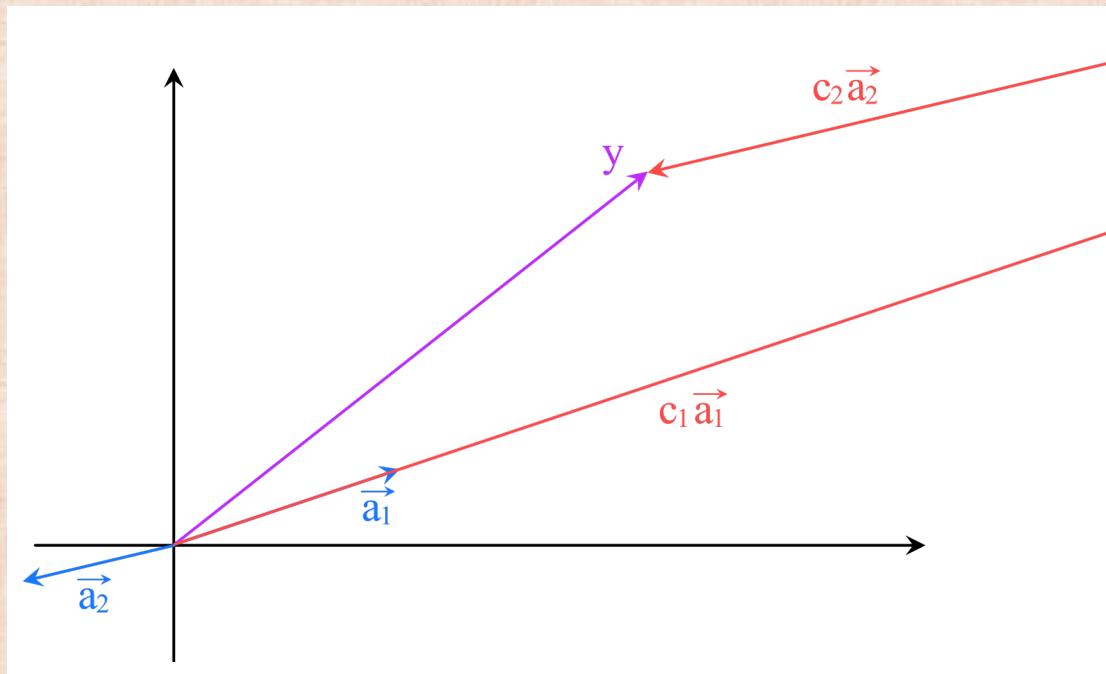
Matrix Columns as Vectors

- Let the k -th column of A be vector a_k , so $Ac = y$ is equivalent to $\sum_k c_k a_k = y$
- That is, find a linear combination of the columns of A that give the right hand side vector y



Matrix Columns as Vectors

- As columns become more parallel, the values of c tend to become arbitrarily large, ill-conditioned, and thus erroneous



- In this example, the column vectors go to far to the right and back in order to (fully) illustrate

Singular Matrices

- If two columns of a matrix are parallel, they may be combined in an infinite number of ways while still obtaining the same result
 - Thus, the problem does not have a unique solution
- In addition, the n columns of A span at most an $n - 1$ dimensional subspace
 - the range of A is at most $n - 1$ dimensional
- If the right hand side vector is not contained in this $n - 1$ dimensional subspace, then the problem has no solution
 - otherwise, there are infinite solutions

Singular Matrices

- If any column of a matrix is a linear combination of other columns, they may be combined in an infinite number of ways while still obtaining the same result
 - Thus, the problem does not have a unique solution
- In addition, the n columns of A span at most an $n - 1$ dimensional subspace
 - the range of A is at most $n - 1$ dimensional
- If the right hand side vector is not contained in this $n - 1$ dimensional subspace, then the problem has no solution
 - otherwise, there are infinite solutions

Near Singular Matrices

- Computational approaches struggle to obtain accuracy, when columns aren't orthogonal enough
- That is, invertible matrices may not be computationally invertible
- We use the concept of a condition number to describe how hard or easy it is to solve a problem computationally

Approximation Errors

- **Modeling errors** – Parts of a problem under consideration may be ignored. For example, when simulating solids/fluids, sometimes frictional/viscous effects are not included.
- **Empirical constants** – Some numbers are unknown and measured in a laboratory only to limited precision. Others may be known more accurately, but limited precision hinders the ability to express them on a finite precision computer. Examples include Avogadro's number, the speed of light in a vacuum, the charge on an electron, Planck's constant, Boltzmann's constant, pi, etc. Note that the speed of light is 299792458 m/s exactly, so we are ok for double precision but not single precision.

Approximation Errors

- **Rounding Errors:** Even integer calculations lead to floating point numbers, e.g. $5/2=2.5$. And floating point calculations frequently admit rounding errors, e.g. $1./3.=.3333333\dots$ cannot be expressed on the computer; thus, the computer commits rounding errors to express numbers with machine precision, e.g. $1./3.=.3333333$. Machine precision is 10^{-7} for single precision and 10^{-16} for double precision.
- **Truncation errors** – Also called discretization errors. These occur in the mathematical approximation of an equation as opposed to the mathematical approximation of the physics (modeling errors). One (typically) cannot take a derivative or integral exactly on the computer, so they are approximated with some formula (recall Simpson's rule from Calculus).

Approximation Errors

- **Inaccurate inputs** – Often, one is only concerned with part of a calculation, and a given set of input numbers is used to produce a set of output numbers. Those inputs may have previously been subjected to any of the errors listed above and thus may already have limited accuracy. This has implications for various algorithms. If inputs are only accurate to 4 decimal places, it makes little sense to carry out an algorithm to an accuracy of 8 decimal places.

Computational Approach

- **Condition Number:** A problem is ill-conditioned if small changes in the input data lead to large changes in the output. Large condition numbers are bad (sensitive), and small condition numbers are good (insensitive). If the relative changes in the input and the output are identical, the condition number is 1.
 - E.g. Near parallel columns in a matrix lead to a poor condition number!
- **Stability and Accuracy:** For well-conditioned problems, one may attempt to solve them on the computer, and then the terms stability and accuracy come into play.
 - Stability refers to whether or not the algorithm can complete itself in any meaningful way. Unstable algorithms tend to give wildly varying, explosive data that usually lead to NaN's.
 - Stability alone does not indicate that the problem has been solved. One also needs to be concerned with the size of the error, which could still be enormous (e.g. no significant digits correct). Accuracy refers to how close an answer is to the correct solution.

Computational Approach

- A problem should be well-posed before even considering it computationally
- Computational Approach:
 - 1) Conditioning - formulate a well-conditioned approach
 - 2) Stability - devise a stable algorithm
 - 3) Accuracy - make the algorithm as accurate as is practical

Vector Norms (Carefully)

- Consider the norm of a vector: $\|x\|_2 = \sqrt{x_1^2 + \cdots + x_m^2}$
- Straightforward algorithm:

```
{ for (i=1,m) sum+=x(i)*x(i); return sqrt(sum); }
```
- This can overflow MAX_FLOAT/MAX_DOUBLE for large m
- Safer algorithm:

```
find z=max(abs(x(i)))  
{ for (i=1,m) sum+=sqr(x(i)/z); return z*sqrt(sum); }
```

Quadratic Formula (Carefully)

- Consider $.0501x^2 - 98.78x + 5.015 = 0$
 - To 10 digits of accuracy: $x \approx 1971.605916$ and $x \approx .05077069387$

- Using 4 digits of accuracy in the quadratic formula gives:

$$\frac{98.78+98.77}{.1002} = 1972 \quad \text{and} \quad \frac{98.78-98.77}{.1002} = .0998$$

- The second root is completely wrong in the leading significant digit!

- De-rationalize: $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ to $\frac{2c}{-b \mp \sqrt{b^2 - 4ac}}$

- Using 4 digits of accuracy in the (second) de-rationalized quadratic formula gives:

$$\frac{10.03}{98.78 - 98.77} = 1003 \quad \text{and} \quad \frac{10.03}{98.78 + 98.77} = .05077$$

- Now the second root is fine, but the first is wrong!

- Conclusion: use one formula for each root

Quadratic Formula (Carefully)

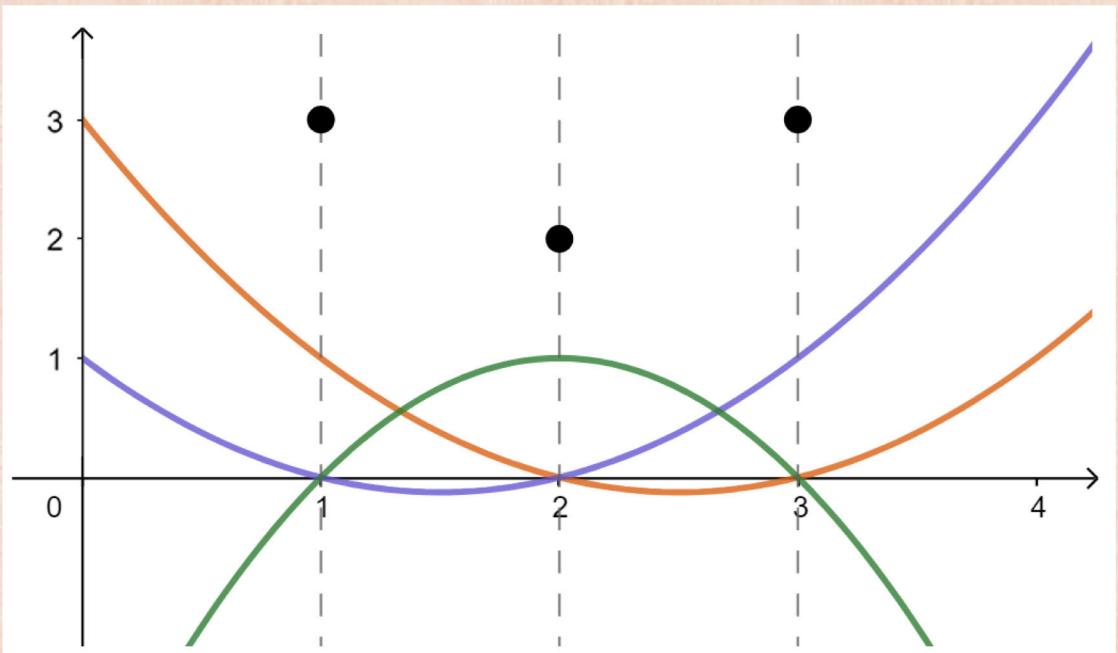
- *Did you know this was an issue?*
- Imagine debugging code with the correct quadratic formula and getting zero digits of accuracy on a test case!
- The specific sequence of operations performed in solving the quadratic formula can result in large errors. Subtractions followed by divisions cause errors. Subtraction reveals the error that round-off makes. Division can amplify round-off error
- It is important to understand that the operations themselves are not dangerous, but the specific order [aka the algorithm] can be

Polynomial Interpolation (Carefully)

- Given basis functions ϕ and unknowns c : $y = c_1\phi_1 + c_2\phi_2 + \cdots + c_n\phi_n$
- Monomial basis: $\phi_k(x) = x^{k-1}$
 - Vandermonde matrix may become near-singular and difficult to invert
- Lagrange Basis: $\phi_k(x) = \frac{\prod_{i \neq k} x - x_i}{\prod_{i \neq k} x_k - x_i}$ so $\phi_k(x_k) = 1$ and $\phi_k(x_i) = 0$ for $i \neq k$
- Write an equation for each point, note that the equations are linear, and put into matrix form (as usual)
- Obtain $Ac = y$ where A is the identity matrix (i.e. $Ic = y$), so $c = y$ trivially
- Evaluation of the polynomial is expensive (lots of terms)
 - i.e. network inference would be expensive

Lagrange Basis for Polynomial Interpolation

- Consider data $(1,3), (2,2), (3,3)$ with quadratic basis functions that are 1 at their corresponding data point and 0 at the other data points



- $\phi_1(x) = \frac{(x-2)(x-3)}{(1-2)(1-3)} = \frac{1}{2}(x-2)(x-3)$
- $\phi_1(1) = 1, \phi_1(2) = 0, \phi_1(3) = 0$
- $\phi_2(x) = \frac{(x-1)(x-3)}{(2-1)(2-3)} = -(x-1)(x-3)$
- $\phi_2(1) = 0, \phi_2(2) = 1, \phi_2(3) = 0$
- $\phi_3(x) = \frac{(x-1)(x-2)}{(3-1)(3-2)} = \frac{1}{2}(x-1)(x-2)$
- $\phi_3(1) = 0, \phi_3(2) = 0, \phi_3(3) = 1$

Newton Basis for Polynomial Interpolation

- Basis functions: $\phi_k(x) = \prod_{i=1}^{k-1} x - x_i$
- Here $Ac = y$ has lower triangular A (as opposed to dense/diagonal)
- Columns don't overlap, and not too expensive to evaluate/inference
- Can solve via a divided difference table:
 - Initially: $f[x_i] = y_i$
 - Then, at each level, recursively: $f[x_1, x_2, \dots, x_k] = \frac{f[x_2, x_3, \dots, x_k] - f[x_1, x_2, \dots, x_{k-1}]}{x_k - x_1}$
 - Finally: $c_k = f[x_1, x_2, \dots, x_k]$
- As usual, high order polynomials still tend to be oscillatory
 - Using unequally spaced data points can help, e.g. Chebyshev points

Summary: Polynomial Interpolation

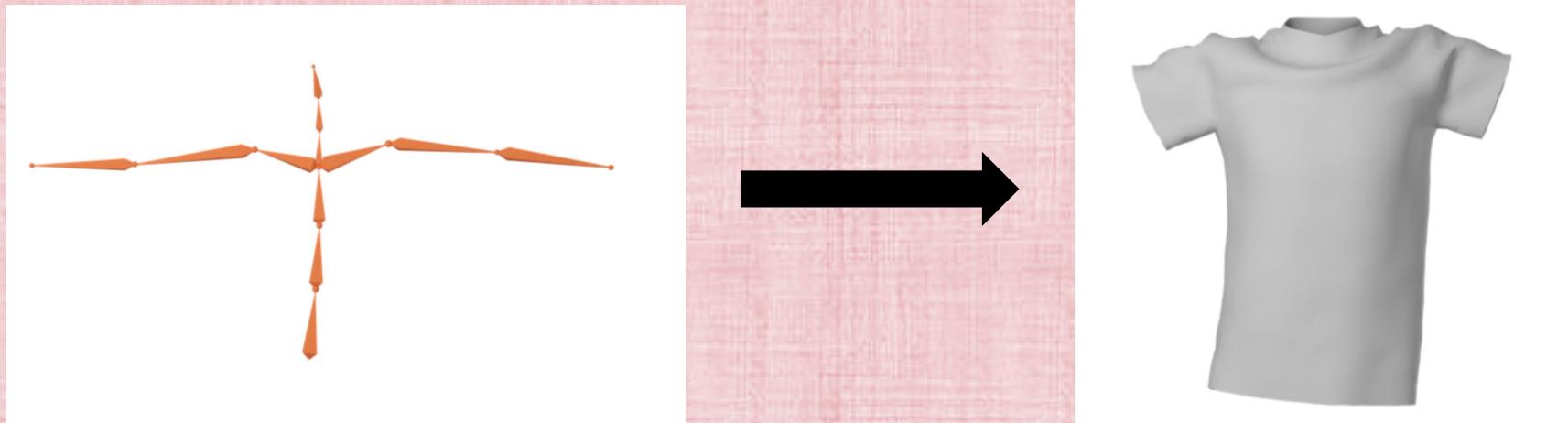
- Monomial/Lagrange/Newton basis all give the same exact unique polynomial
 - as one can see by multiplying out and collecting like terms
- But the representation makes finding/evaluating the polynomial easier/harder

Representation Matters

- Consider: Divide CCX by VI
- As compared to: Divide 210 by 6
- See Chapter 15 on Representation Learning in the Deep Learning book

Predict 3D Cloth Shape from Body Pose (Carefully)

- Input: pose parameters θ are joint rotation matrices
 - 10 upper body joints with a 3×3 rotation matrix for each gives a $90D$ pose vector (should be $30D$ using quaternions)
 - global translation/rotation of root frame is ignored
- Output: $3D$ cloth shape φ
 - 3,000 vertices in a cloth triangle mesh gives a $9,000D$ shape vector
- Function $f: \mathbf{R}^{90} \rightarrow \mathbf{R}^{9000}$

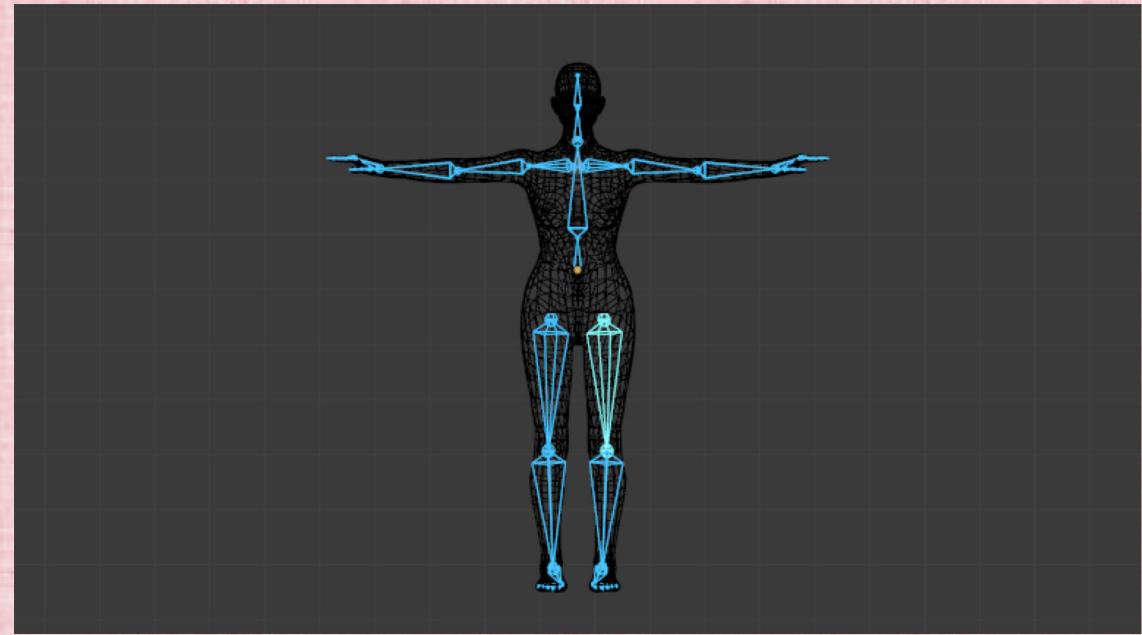


Approach

- Given: m training data points (θ_i, φ_i) generated from the true/approximated function $\varphi_i = f(\theta_i)$
- E.g. using simulation or capture
- Goal: learn an \hat{f} that approximates f , i.e. $\hat{f}(\theta) = \hat{\varphi} \approx \varphi = f(\theta)$
- Issue: As joints rotate (rotation is highly nonlinear), cloth vertices move in complex nonlinear ways that are difficult to capture with a network, i.e. in \hat{f}
- How should the nonlinear rotations be handled?

Aside: Procedural Skinning

- Deforms a body surface mesh to match a skeletal pose
 - well studied and widely used in graphics
- In the rest pose, associate each vertex of the body surface mesh with a few joints/bones
- A weight from each joint/bone dictates how much impact its pose has on the vertex's position
- As the pose changes, joint/bone changes dictate new positions for skin vertices



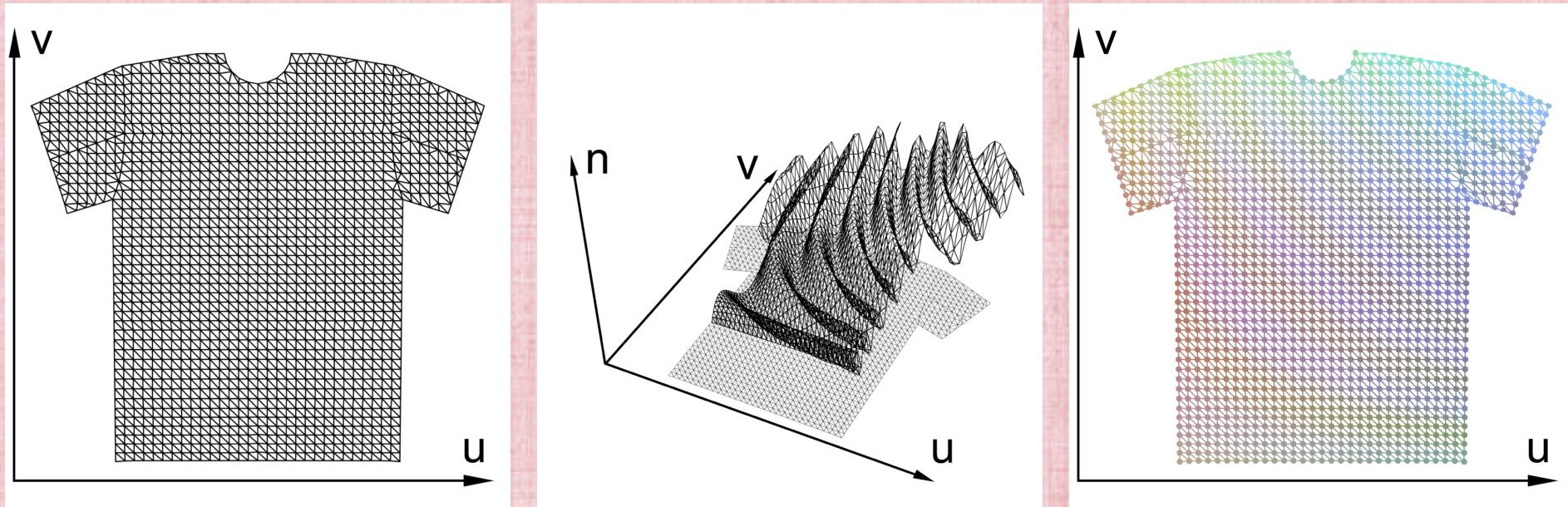
Picture from Blender website [link](#)

Leverage Procedural Skinning

- Leverage the plethora of prior work on procedural skinning to estimate the body surface mesh based on pose parameters, $S(\theta)$
- Then, represent the cloth mesh as offsets from the skinned body mesh, $D(\theta)$
- Overall, $\hat{\varphi}(\theta) = S(\theta) + D(\theta)$, where only $D(\theta)$ needs to be learned
- The procedural skinning prior $S(\theta)$ captures much of the nonlinearities, so that the remaining $D(\theta)$ is a smoother function and thus easier to approximate/learn

Pixel Based Cloth

- Assign texture coordinates to a cloth triangle mesh
- Then, transfer the mesh into pattern/texture space (left)
- Store (u, v, n) offsets in the pattern/texture space (middle)
- Convert (u, v, n) offsets to RGB color values or “pixels” (right)



Body Skinning of Cloth Pixels

- Shrink-wrap the cloth pixels (left) to the body triangle mesh (middle)
 - barycentrically embed cloth pixels to follow body mesh triangles
- As the body deforms, cloth pixels move with their parent triangles (right)
- Then, as a function of pose θ , learn per-pixel (u, v, n) offsets $D(\theta)$ from the skinned cloth pixels $S(\theta)$ to the actual cloth mesh φ



RGB Values of Cloth Pixels Correspond to Offset Vectors

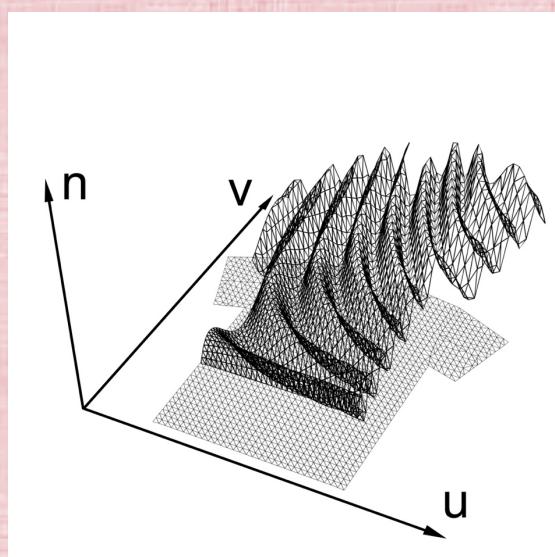
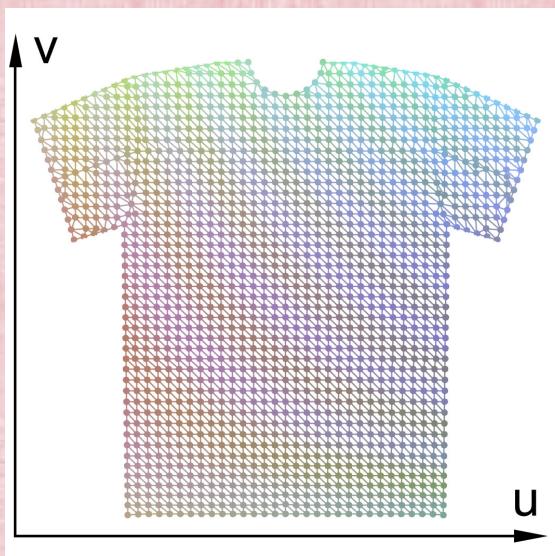
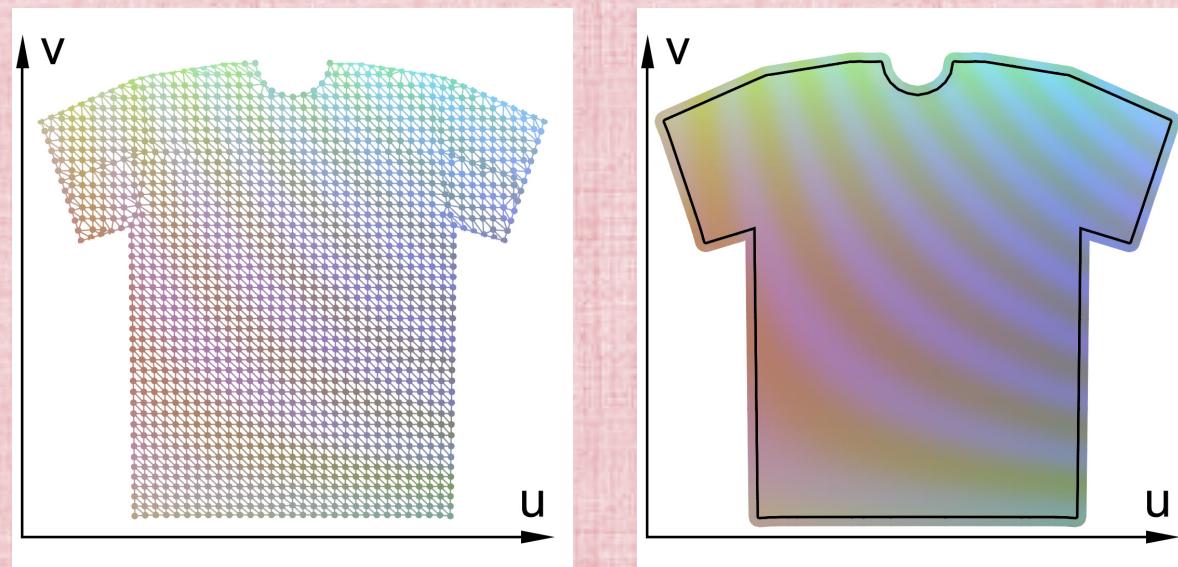


Image Based Cloth

- Rasterize triangle vertex colors to standard 2D image pixels (in pattern space)
- Function output becomes a (standard) 2D RGB image
- More continuous than the cloth pixels (which have discrete topology)
- Now, can learn with standard Convolutional Neural Network (CNN) techniques



Encode 3D Cloth Shapes as 2D Images

- For each pose in the training data, calculate per-vertex offsets and rasterize them into an image in pattern space
- Then learn to predict an image from pose parameters, $I(\theta)$
- Given $I(\theta)$, interpolate to vertex positions (cloth pixels) and convert to offsets that are added to the skinned vertex positions: $\hat{\varphi}(\theta) = S(\theta) + h(I(\theta))$



Unit 2

Linear Systems

Motivation

- “Matrices are bad, vector spaces are good”
 - That is, don’t think of matrices as a collection of numbers
 - Rather, think of columns as vectors in a high dimensional space
- We don’t have great intuition going from R^1 to R^2 to R^3 to R^n (for large n)
- Thus, we consider vectors in high dimensional spaces as a way of gaining intuition about what’s going on
- Linear algebra, as a mathematical area, contains a lot of machinery for dealing with, discussing, and gaining intuition about vectors in high dimensional spaces
- So, while we will cover the essentials of linear algebra, we will do it from the standpoint of understanding higher dimensional spaces

System of Linear Equations

- System of equations: $3c_1 + 2c_2 = 6$ and $-4c_1 + c_2 = 7$
- Matrix form: $\begin{pmatrix} 3 & 2 \\ -4 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 6 \\ 7 \end{pmatrix}$ or $Ac = b$
- Given A and b , determine c
- There is a unique solution, no solution, or infinite solutions
- Ideally, software would determine whether there was a unique solution, no solution, or infinite solutions. In the last case, it should list the parameterized family of solutions. Unfortunately, this turns out to be difficult
- Note: in this class, x will be typically be used for **data**, and c will typically be used for **unknowns** (such as for the unknown parameters of a neural network)

“Zero”

- One of the basic issues that has to be confronted is the concept of “zero”
- When dealing with large numbers, e.g. Avogadro’s number, $6.022e23$, zero can be quite large
 - E.g. $6.022e23 - 1e7 = 6.022e23$ in double precision, making $1e7$ behave like “zero”
- When dealing with small numbers, such as $1e-23$, zero is much smaller
 - In this case, on the order of $1e-39$ in double precision
- Mixing big and small numbers often wreaks havoc on algorithms
- So we typically non-dimensionalize and normalize to make equations $O(1)$ as opposed to $O(\text{"big"})$ or $O(\text{"small"})$

Row/Column Scaling

- Consider:
$$\begin{pmatrix} 3e6 & 2e10 \\ 1e-4 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 5e10 \\ 6 \end{pmatrix}$$
- Row Scaling - divide first row by $1e10$ to obtain:
$$\begin{pmatrix} 3e-4 & 2 \\ 1e-4 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \end{pmatrix}$$
- Column Scaling - define a new variable $c_3 = (1e-4)c_1$ to obtain:
$$\begin{pmatrix} 3 & 2 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c_3 \\ c_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 6 \end{pmatrix}$$
- Final result is much easier to treat with finite precision arithmetic
- Solve for c_3 and c_2 , and then $c_1 = (1e4)c_3$

Transpose and Symmetry

- Elements of a matrix are often referred to by their row and column
- For example, a_{ik} is the element of matrix A in row i and column k
- Transpose swaps all the row and column of every entry
- A^T moves element a_{ik} to row k column i
- The size of the matrix will change if it is rectangular: $\begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}^T = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$

Symmetric Matrices have $A^T = A$ meaning that $a_{ik} = a_{ki}$ for all i and k

Square Matrix

- A size mxn matrix has m rows and n columns
- For now, let's consider square nxn matrices
- We will consider rectangular matrices with $m \neq n$ later

Solvability

- Singular – A is singular when it is not invertible (does not have an inverse)
- Various ways of showing this:
 - A column is linearly dependent on others (as we have seen before)
 - The determinant is zero: $\det A = 0$
 - A has a nonempty null space, i.e. $\exists c \neq 0$ with $Ac = 0$
- Rank - maximum number of linearly independent columns
- Singular matrices have rank $< n$ (the # of columns), i.e. rank-deficient, and have either no solution or infinite solutions
- A nonsingular square matrix has an inverse: $AA^{-1} = A^{-1}A = I$
 - so $Ac = b$ can be solved for c via $c = A^{-1}b$
- *Note: we typically do not compute the inverse, but instead have a solution algorithm that exploits its existence*

Matrices as Vectors (an example)

- Recall $Ac = \sum_k c_k a_k$ where the a_k are the columns of A
- Consider $Ac = 0$ or $\sum_k c_k a_k = 0$
- If one column is a linear combination of others, then the linear combination weights can be used to obtain $Ac = 0$ with c nonzero
 - This nonzero c is in the null space of A , and A is singular
- Conversely, if the only solution to $Ac = 0$ is c identically 0, then no column is linearly dependent on the others
 - Thus A is nonsingular

Diagonal Matrices

- All off-diagonal entries are 0
- Equations are decoupled, and easy to solve
- E.g. $\begin{pmatrix} 5 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 10 \\ -1 \end{pmatrix}$ has $5c_1 = 10$ and $2c_2 = -1$ so $c_1 = 2$ and $c_2 = -.5$
- A zero on the diagonal indicates a singular system, which has no solution, e.g. $0c_1 = 10$, or infinite solutions, e.g. $0c_1 = 0$
- The determinant of a diagonal matrix is obtained by multiplying all the diagonal elements together
- Thus, a 0 on the diagonal implies a zero determinant and a singular matrix

Upper Triangular Matrices

- All entries below the diagonal are 0
- Nonsingular when the diagonal elements are all nonzero
 - Determinant is obtained by multiplying all the diagonal elements together
- Solved via back substitution

- E.g. consider $\begin{pmatrix} 5 & 3 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 10 \\ 10 \end{pmatrix}$
- Start at the bottom with $5c_3 = 10$ or $c_3 = 2$, and move upwards one row at a time. Next, $c_2 - c_3 = 10$ or $c_2 - 2 = 10$ or $c_2 = 12$. Then, $5c_1 + 3c_2 + c_3 = 0$ or $5c_1 + 36 + 2 = 0$ or $c_1 = -38/5 = -7.6$

Lower Triangular Matrices

- All entries above the diagonal are 0
- Nonsingular when the diagonal elements are all nonzero
 - Determinant is obtained by multiplying all the diagonal elements together
- Solved via forward substitution

- E.g. consider $\begin{pmatrix} 5 & 0 & 0 \\ -1 & 1 & 0 \\ 1 & 3 & 5 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 0 \end{pmatrix}$
- Start at the top with $5c_1 = 10$ or $c_1 = 2$, and move downwards one row at a time. Next, $-c_1 + c_2 = 10$ or $-2 + c_2 = 10$ or $c_2 = 12$. Then, $c_1 + 3c_2 + 5c_3 = 0$ or $2 + 36 + 5c_3 = 0$ or $c_3 = -38/5 = -7.6$

Elimination Matrix

- Standard basis vectors: $\hat{e}_i = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ with a 1 in the i -th row/entry
- Given $\begin{pmatrix} a_{1k} \\ \vdots \\ a_{ik} \\ a_{i+1,k} \\ \vdots \\ a_{mk} \end{pmatrix}$, define $m_{ik} = \frac{1}{a_{ik}} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ a_{i+1,k} \\ \vdots \\ a_{mk} \end{pmatrix}$ and $M_{ik} = I_{mxm} - m_{ik}\hat{e}_i^T$
- M_{ik} is a size mxm elimination matrix that subtracts multiples of row i from rows $> i$ in order to create zeroes in column k

Elimination Matrix

- Let $a_1 = \begin{pmatrix} 2 \\ 4 \\ -2 \end{pmatrix}$
- $M_{11} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 0 \\ 4 \\ -2 \end{pmatrix} (1 \quad 0 \quad 0) = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ and $M_{11}a_1 = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix}$
- $M_{21} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{4} \begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix} (0 \quad 1 \quad 0) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1/2 & 1 \end{pmatrix}$ and $M_{21}a_1 = \begin{pmatrix} 2 \\ 4 \\ 0 \end{pmatrix}$

Elimination Matrix Inverse

- Inverse of an elimination matrix is $L_{ik} = M_{ik}^{-1} = I_{m \times m} + m_{ik} \hat{e}_i^T$
- L_{ik} is a size $m \times m$ elimination matrix that adds multiples of row i to rows $> i$ in order to reverse the effect of M_{ik}

- $L_{11} = M_{11}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$

- $L_{21} = M_{21}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1/2 & 1 \end{pmatrix}$

Combining Elimination Matrices

- $M_{i_1 k_1} M_{i_2 k_2} = I - m_{i_1 k_1} \hat{e}_{i_1}^T - m_{i_2 k_2} \hat{e}_{i_2}^T$ when $i_1 < i_2$ but not when $i_1 > i_2$

$$M_{11} M_{21} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 1/2 & 1 \end{pmatrix}$$

- $L_{i_1 k_1} L_{i_2 k_2} = I + m_{i_1 k_1} \hat{e}_{i_1}^T + m_{i_2 k_2} \hat{e}_{i_2}^T$ when $i_1 < i_2$ but not when $i_1 > i_2$

$$L_{11} L_{21} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & -1/2 & 1 \end{pmatrix}$$

Gaussian Elimination

- Consider $\begin{pmatrix} 2 & 4 & -2 \\ 4 & 9 & -3 \\ -2 & -3 & 7 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 8 \\ 10 \end{pmatrix}$
- $M_{11}A = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 4 & -2 \\ 4 & 9 & -3 \\ -2 & -3 & 7 \end{pmatrix} = \begin{pmatrix} 2 & 4 & -2 \\ 0 & 1 & 1 \\ 0 & 1 & 5 \end{pmatrix}$ and $M_{11}b = \begin{pmatrix} 2 \\ 4 \\ 12 \end{pmatrix}$
- $M_{22}M_{11}A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 4 & -2 \\ 0 & 1 & 1 \\ 0 & 1 & 5 \end{pmatrix} = \begin{pmatrix} 2 & 4 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{pmatrix}$ and $M_{22}M_{11}b = \begin{pmatrix} 2 \\ 4 \\ 8 \end{pmatrix}$
- Then, solve $\begin{pmatrix} 2 & 4 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \\ 8 \end{pmatrix}$ via back substitution

LU Factorization

- Gaussian Elimination gives an upper triangular $U = M_{n-1,n-1} \cdots M_{22}M_{11}A$
- Using inverses, $A = L_{11}L_{22} \cdots L_{n-1,n-1}M_{n-1,n-1} \cdots M_{22}M_{11}A = L_{11}L_{22} \cdots L_{n-1,n-1}U$
- Since $L_{i_1 i_1} L_{i_2 i_2} = I + m_{i_1 i_1} \hat{e}_{i_1}^T + m_{i_2 i_2} \hat{e}_{i_2}^T$ when $i_1 < i_2$, $L = L_{11}L_{22} \cdots L_{n-1,n-1}$ is lower triangular and $A = LU$

- Here $L = L_{11}L_{22} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix}$

$$A = \begin{pmatrix} 2 & 4 & -2 \\ 4 & 9 & -3 \\ -2 & -3 & 7 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 4 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & 4 \end{pmatrix} = LU$$

LU Factorization

- Factorizing $A = LU$ facilitates solving $Ac = b$
- In order to solve $LUc = b$, define an auxiliary variable $\hat{c} = Uc$
- First, solve $L\hat{c} = b$ for \hat{c} via forward substitution
- Second, solve $Uc = \hat{c}$ for c via back substitution
- Note: the LU factorization is only computed once, and then can be used afterwards on many right hand side (b) vectors

Pivoting

- $A = \begin{pmatrix} 0 & 4 \\ 4 & 9 \end{pmatrix}$ requires division by zero when calculating M_{11}
- (Partial) Pivoting - swap rows to use the largest (magnitude) element in the column under consideration
 - Don't forget to swap the right hand side b too
- Full Pivoting swap rows and columns to use the largest possible element
 - Don't forget to change the order of the unknowns c
- When considering column k , can only swap with rows/columns $\geq k$

Permutation Matrix

- Constructed by switching the 2 rows of I that one wants swapped
- E.g. $P_{13} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$; $P_{13}A$ swaps the first and third rows of A
- Permutation matrices are their own inverses (swapping again restores the rows)
- Switching rows i_1 and i_2 moves a 1 from $a_{i_1 i_1}$ to $a_{i_2 i_1}$ and from $a_{i_2 i_2}$ to $a_{i_1 i_2}$ preserving symmetry (i.e. $P_{i_1 i_2}^T = P_{i_1 i_2}$)
- To swap the first and third unknowns: $Ac = AP_{13}P_{13}c = (AP_{13})(P_{13}c)$ where $P_{13}c$ has the unknowns swapped and AP_{13} swaps the columns

Full Pivoting

- Let P_{r_i} be the permutation matrix that (potentially) switches row i with a row $> i$
- Let P_{c_k} be the permutation matrix that (potentially) switches column k with a col $> k$
- Then full pivoting can be written as:
$$(M_{n-1,n-1}P_{r_{n-1}} \cdots M_{22}P_{r_2}M_{11}P_{r_1}AP_{c_1}P_{c_2} \cdots P_{c_{n-1}})(P_{c_{n-1}} \cdots P_{c_2}P_{c_1}c)$$
- Once known, $P_r = P_{r_{n-1}} \cdots P_{r_2}P_{r_1}$ and $P_c = P_{c_{n-1}} \cdots P_{c_2}P_{c_1}$ can be used to do all the permutations first so that the result doesn't require pivoting
- $Ac = b$ becomes $(P_rAP_c^T)(P_c c) = P_r b$ or $A_P c_P = b_P$; then, $A_P = L_P U_P$ can be computed without pivoting
- Subsequently, given right hand side b , solve $L_P U_P c_P = P_r b$ to find c_P using forward/back substitution, and then $c = P_c^T c_P$

Sparsity

- Most large matrices (of interest) operate on variables that only interact with a sparse set of other variables
- This makes the matrix sparse (as opposed to dense), i.e. most entries are 0
- However, the inverse of a sparse matrix can contain an unwieldy amount of non-zero entries
- E.g. the Poisson equation on a relatively small 100^3 Cartesian grid has an unknown for each of the 10^6 grid points
- For each unknown, the discretized Poisson equation depends on the unknown itself and its 6 immediate neighbors
- Thus, the size $10^6 \times 10^6$ matrix has only 7×10^6 nonzero entries
- But the inverse potentially has 10^{12} nonzero entries

Computing the Inverse

- When A is relatively small (and dense), one might compute A^{-1}
- Since $AA^{-1} = I$, the solution c_k to $Ac_k = \hat{e}_k$ is the k-th column of A^{-1}
- First, compute $A_P = L_P U_P$ as usual
- Then, solve $Ac_k = \hat{e}_k$ repeatedly (n times, once for each column)

Unit 3

Understanding Matrices

Eigensystem

- Eigenvalues - special directions v_k in which a matrix only applies scaling
- Eigenvalues - the amount λ_k of that scaling
- Left Eigenvectors (or simply eigenvectors) satisfy $A v_k = \lambda_k v_k$
 - Eigenvectors represent directions, so $A(\alpha v_k) = \lambda_k(\alpha v_k)$ is also true for all α
- Right Eigenvectors satisfy $u_k^T A = \lambda_k u_k^T$ (or $A^T u_k = \lambda_k u_k$)
- Diagonal matrices have eigenvalues on the diagonal, and eigenvectors \hat{e}_k

$$\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = 3 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

- Upper/lower triangular matrices also have eigenvalues on the diagonal

$$\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3 \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Complex Numbers

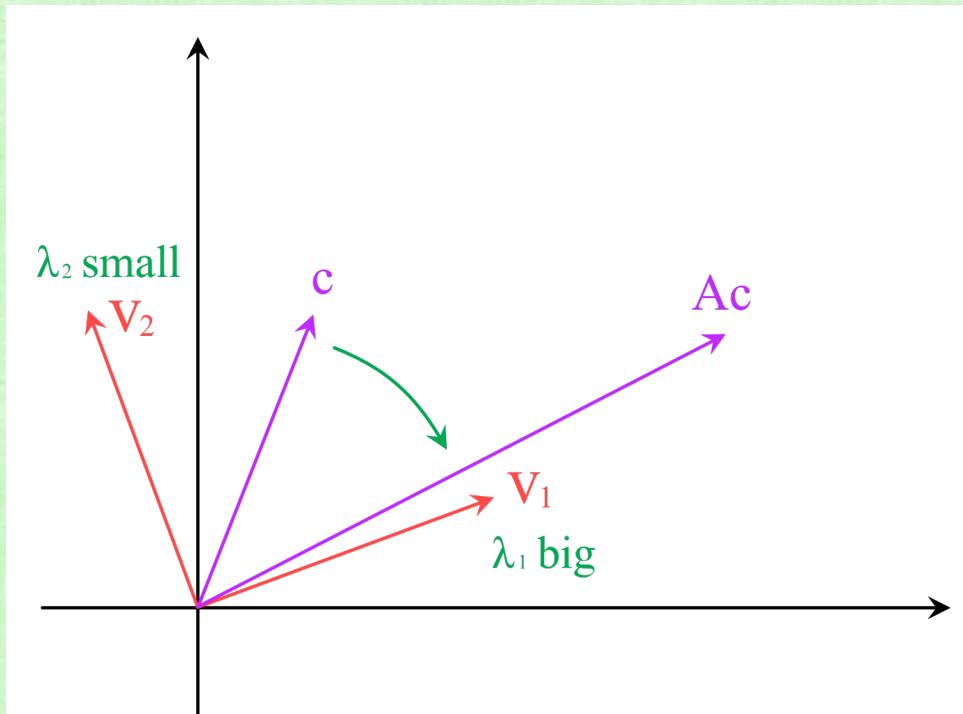
- Complex numbers may appear in both eigenvalues and eigenvectors

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ i \end{pmatrix} = i \begin{pmatrix} 1 \\ i \end{pmatrix}$$

- Recall: complex conjugate: $(a + bi)^* = a - bi$
- Hermitian Matrix: $A^{*T} = A$ (often A^{*T} is written as A^H)
 - $A\boldsymbol{\nu} = \lambda\boldsymbol{\nu}$ implies $(A\boldsymbol{\nu})^{*T} = (\lambda\boldsymbol{\nu})^{*T}$ or $\boldsymbol{\nu}^{*T}A = \lambda^*\boldsymbol{\nu}^{*T}$
 - Using this, $A\boldsymbol{\nu} = \lambda\boldsymbol{\nu}$ implies $\boldsymbol{\nu}^{*T}A\boldsymbol{\nu} = \lambda\boldsymbol{\nu}^{*T}\boldsymbol{\nu}$ or $\lambda^*\boldsymbol{\nu}^{*T}\boldsymbol{\nu} = \lambda\boldsymbol{\nu}^{*T}\boldsymbol{\nu}$ or $\lambda^* = \lambda$
 - Thus, Hermitian matrices have $\lambda \in \mathbb{R}$ (no complex eigenvalues)
- Symmetric real-valued matrices have real-valued eigenvalues/eigenvectors
 - However, complex eigenvectors work too, e.g. $A(\alpha\boldsymbol{\nu}_k) = \lambda_k(\alpha\boldsymbol{\nu}_k)$ with α complex

Spatial Deformation

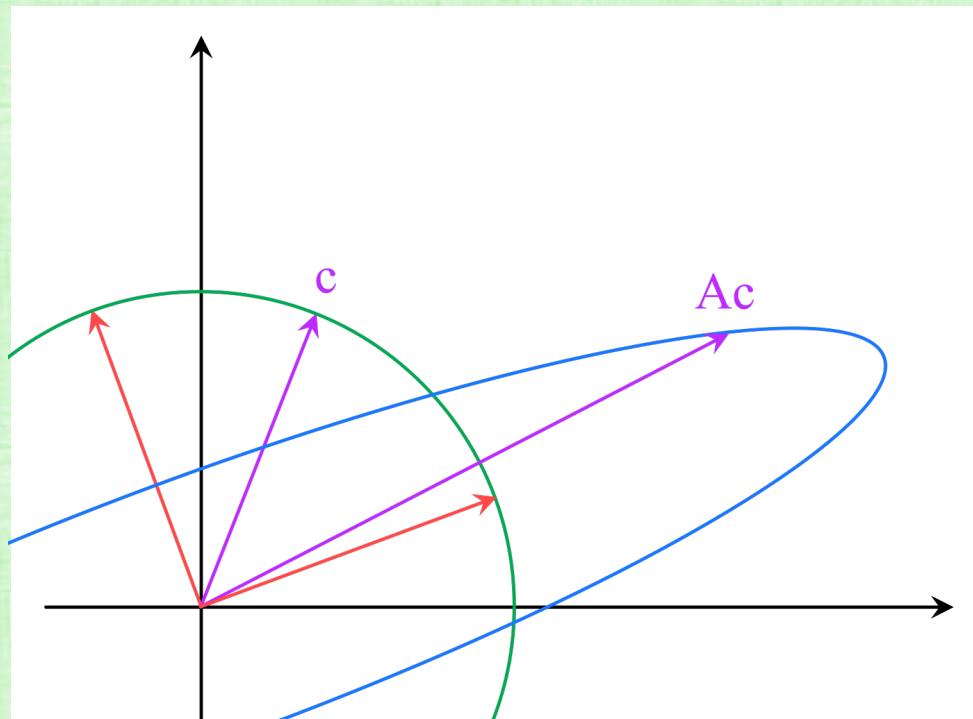
- Suppose $c = \sum_k \alpha_k v_k$ so that $Ac = \sum_k \alpha_k A v_k = \sum_k \alpha_k \lambda_k v_k$
- Thus, A tilts c away from directions with smaller eigenvalues and towards directions with larger eigenvalues



- Large λ_k stretch components in their associated v_k directions
- Small λ_k squish components in their associated v_k directions
- Negative λ_k flip the sign of components in their associated v_k directions

Spatial Deformation

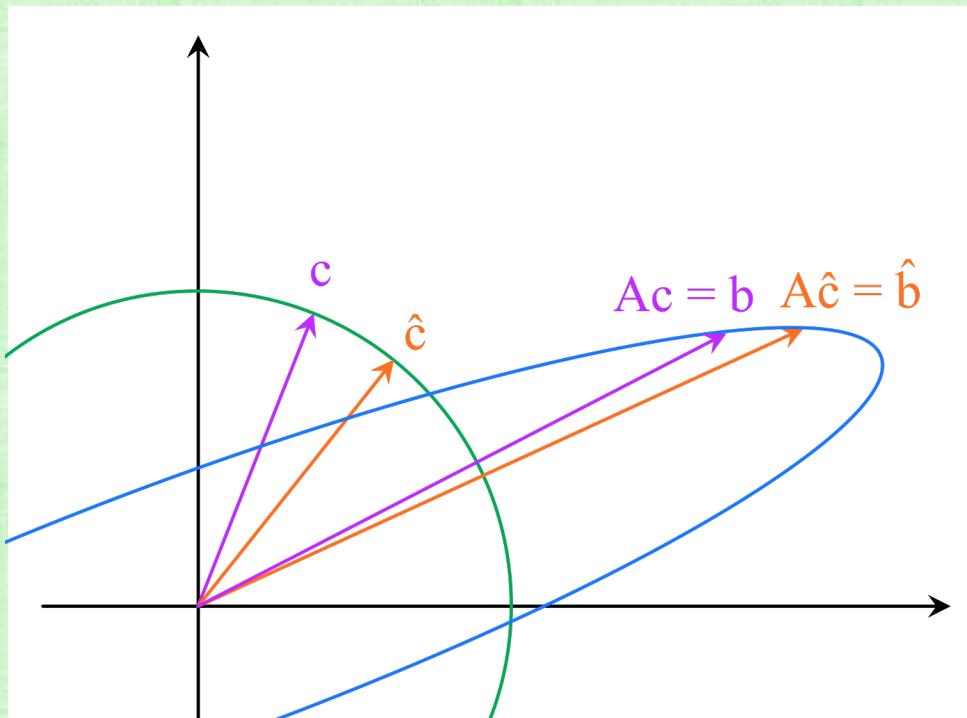
- Consider every point on a unit circle as a vector $c = \sum_k \alpha_k v_k$, and remap each point via $Ac = \sum_k \alpha_k \lambda_k v_k$



- The remapped shape (blue) is more elliptical than the original circle (green)
- The circle is stretched along the axis with the larger eigenvalue, and compressed along the axis with the smaller eigenvalue
- The larger the ratio of eigenvalues, the more elliptical the circle becomes

Solving Linear Systems

- Perturb the right hand side from b to \hat{b} , and solve $A\hat{c} = \hat{b}$ to find \hat{c}
- Note how c and \hat{c} are more separated than b and \hat{b} , i.e. the solution is more perturbed than the right hand side



- Small right hand side changes lead to larger changes in the solution
- **Small algorithmic errors are also amplified**, since they behave similar to small right hand side changes
- The amount of amplification is proportional to the ratio of the eigenvalues

Preconditioning

- Suppose A has very big eigenvalue ratios, making $Ac = b$ difficult to solve
- Suppose one had an approximate guess for the inverse, i.e an $\hat{A}^{-1} \approx A^{-1}$
- Then, transform $Ac = b$ into $\hat{A}^{-1}Ac = \hat{A}^{-1}b$ or $\hat{I}c = \tilde{b}$
 - Typically, a bit more involved than this, but conceptually the same
- \hat{I} is not the identity, so there is still more work to do in order to find c
- However, \hat{I} has similar size eigenvalues (clusters work too), making $\hat{I}c = \tilde{b}$ far easier to solve than a poorly conditioned $Ac = b$

Preconditioning works GREAT!

- It is best to re-scale ellipsoids along eigenvector axes, but scaling along the coordinate axes (diagonal/Jacobi preconditioning) works well too

Rectangular Matrices (Rank)

- An $m \times n$ rectangular matrix has m rows and n columns
- (Note: these comments also hold for square matrices with $m = n$)
- The columns span a space, and the unknowns are weights on each column (recall $\mathbf{A}\mathbf{c} = \sum_k c_k \mathbf{a}_k$)
- A matrix with n columns has maximum rank n
- The actual rank depends on how many of the columns are linearly independent from one another
- Each column has length m (which happens to be the number of rows)
- Thus, the columns live in m spatial dimensions, and at best can span that whole space
- That is, there is a maximum of m independent columns
- Overall, a matrix at most has rank equal to the minimum of m and n
- Both considerations are based on looking at the columns (which are scaled by the unknowns)

Rectangular Matrices (Rank)

- One can find discussions on rows, row spaces, etc. that are used for various purposes
- Although these are fine discussions in regards to matrices/mathematics, they are unnecessary for an intuitive understanding of vector spaces in high dimensions (and as such can be ignored)
- The number of columns is identical to number of variables, which depends on the parameters of the problem
 - E.g. the unknown parameters that govern a neural network architecture
- The number of rows depends on the amount of data used, and adding/removing data does not intrinsically effect the nature of the problem
 - E. g. it does not change the network architecture, but merely perturbs the unknown parameters

Singular Value Decomposition (SVD)

- Factorization of any size $m \times n$ matrix: $A = U\Sigma V^T$
- Σ is $m \times n$ diagonal with non-negative diagonal entries (called singular values)
- U is $m \times m$ orthogonal, V is $n \times n$ orthogonal (their columns are called singular vectors)
 - Orthogonal matrices have orthonormal columns (an orthonormal basis), so their transpose is their inverse. They preserve inner products, and thus are rotations, reflections, and combinations thereof
 - If A has complex entries, then U and V are unitary (conjugate transpose is their inverse)
- Introduced and rediscovered many times: Beltrami 1873, Jordan 1875, Sylvester 1889, Autonne 1913, Eckart and Young 1936. Pearson introduced principle component analysis (PCA) in 1901, which uses SVD. Numerical methods by Chan, Businger, Golub, Kahan, etc.

(Rectangular) Diagonal Matrices

- All off-diagonal entries are 0
 - Diagonal entries are a_{kk} , and off diagonal entries are a_{ki} with $k \neq i$
- E.g. $\begin{pmatrix} 5 & 0 \\ 0 & 2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 10 \\ -1 \\ \alpha \end{pmatrix}$ has $5c_1 = 10$ and $2c_2 = -1$ so $c_1 = 2$ and $c_2 = -0.5$
 - Note that $\alpha \neq 0$ imposes a “no solution” condition
- E.g. $\begin{pmatrix} 5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 10 \\ -1 \\ 0 \end{pmatrix}$ has $5c_1 = 10$ and $2c_2 = -1$ so $c_1 = 2$ and $c_2 = -0.5$
- A zero on the diagonal indicates a singular system, which has no solution, e.g. $0c_1 = 10$, or infinite solutions, e.g. $0c_1 = 0$

Singular Value Decomposition (SVD)

- $A^T A = V \Sigma^T U^T U \Sigma V^T = V (\Sigma^T \Sigma) V^T$, so $(A^T A)v = \lambda v$ is $(\Sigma^T \Sigma)(V^T v) = \lambda(V^T v)$
- $\Sigma^T \Sigma$ is $n \times n$ diagonal with eigenvectors \hat{e}_k , so $\hat{e}_k = V^T v$ and $v = V \hat{e}_k$
- That is, the columns of V are the eigenvectors of $A^T A$
- $AA^T = U \Sigma V^T V \Sigma^T U^T = U (\Sigma \Sigma^T) U^T$, so $(AA^T)v = \lambda v$ is $(\Sigma \Sigma^T)(U^T v) = \lambda(U^T v)$
- $\Sigma \Sigma^T$ is $m \times m$ diagonal with eigenvectors \hat{e}_k , so $\hat{e}_k = U^T v$ and $v = U \hat{e}_k$
- That is, the columns of U are the eigenvectors of AA^T
- When $m \neq n$, either $\Sigma^T \Sigma$ or $\Sigma \Sigma^T$ is larger and contains extra zeros on the diagonal
- Otherwise, their diagonal entries are the squares of the singular values
- That is, the singular values are the (non-negative) square roots of the non-extra eigenvalues of $A^T A$ and AA^T
- Note that both $A^T A$ and AA^T are symmetric positive semi-definite, and thus easy to work with
- E.g. symmetry means their eigensystem (and thus the SVD) has no complex numbers when A doesn't

Example (Tall Matrix)

- Consider size 4×3 matrix $A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix}$
- Label the columns $a_1 = \begin{pmatrix} 1 \\ 4 \\ 7 \\ 10 \end{pmatrix}$, $a_2 = \begin{pmatrix} 2 \\ 5 \\ 8 \\ 11 \end{pmatrix}$, $a_3 = \begin{pmatrix} 3 \\ 6 \\ 9 \\ 12 \end{pmatrix}$
- Since a_1 and a_2 point in different directions, A is at least rank 2
- $a_3 = 2a_2 - a_1$, so the third column is in the span of the first two columns
- Thus A is only rank 2 (not rank 3)

Example (SVD)

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix} =$$

$$\begin{pmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{pmatrix} \begin{pmatrix} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{pmatrix}$$

- Singular values are **25.5**, **1.29**, and **0**
- Singular value of **0** indicates that the matrix is rank deficient
- The rank of a matrix is equal to its number of nonzero singular values

Derivation from $A^T A$ and AA^T

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix} =$$

$$\begin{pmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{pmatrix} \begin{pmatrix} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{pmatrix}$$

- $A^T A$ is size $3x3$ and has 3 eigenvectors (seen in V)
- The square roots of the 3 eigenvalues of $A^T A$ are seen Σ (color coded to the eigenvectors)
- AA^T is size $4x4$ and has 4 eigenvectors (seen in U)
- The square roots of 3 of the eigenvalues of AA^T are seen Σ
 - The 4th eigenvalue of AA^T is an extra eigenvalue of 0

Understanding Ac

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix} =$$
$$\begin{pmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{pmatrix} \begin{pmatrix} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{pmatrix}$$

- A maps from R^3 to R^4
- Ac first projects $c \in R^3$ onto the 3 basis vectors in V
- Then, the associated singular values (diagonally) scale the results
- Lastly, those scaled results are used as weights on the 4 basis vectors in U

Understanding Ac

$$\begin{aligned}
 Ac &= \begin{pmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{pmatrix} \begin{pmatrix} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \\
 &= \begin{pmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{pmatrix} \begin{pmatrix} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v_1^T c \\ v_2^T c \\ v_3^T c \end{pmatrix} \\
 &= \begin{pmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{pmatrix} \begin{pmatrix} \sigma_1 v_1^T c \\ \sigma_2 v_2^T c \\ \sigma_3 v_3^T c \\ 0 \end{pmatrix} \\
 &= u_1 \sigma_1 v_1^T c + u_2 \sigma_2 v_2^T c + u_3 \sigma_3 v_3^T c + u_4 0
 \end{aligned}$$

- Ac projects c onto the basis vectors in V , then scales by the associated singular values, and lastly uses those results as weights on the basis vectors in U

Extra Dimensions

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix} =$$

$$\left(\begin{array}{cccc} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{array} \right) \left(\begin{array}{ccc} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \end{array} \right) \left(\begin{array}{ccc} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{array} \right)$$

- The 3D space of vector inputs can only span a 3D subspace of R^4
- The last (green) column of U represents the unreachable dimension, orthogonal to the range of A , and is always multiplied by 0
- One can delete this column and the associated portion of Σ (and still obtain a valid factorization)

Zero Singular Values

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix} =$$

$$\left(\begin{array}{ccc|cc} .141 & .825 & -.420 & -.351 & \\ .344 & .426 & .298 & .782 & \\ .547 & .028 & .644 & -.509 & \\ .750 & -.371 & -.542 & .079 & \end{array} \right) \left(\begin{array}{cc|c} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right) \left(\begin{array}{ccc} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{array} \right)$$

- The 3rd singular value is 0, so A has a 1D null space that reduces the 3D input vectors to only 2 dimensions
- The associated (pink) terms make no contribution to the final result, and can also be deleted (still obtaining a valid factorization)
- The first 2 columns of U span the 2D subset of R^4 that comprises the range of A

Approximating A

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix} \approx$$

$$\left(\begin{array}{c|cc|cc} .141 & .825 & -.420 & -.351 & \\ \hline .344 & .426 & .298 & .782 & \\ .547 & .028 & .644 & -.509 & \\ .750 & -.371 & -.542 & .079 & \end{array} \right) \left(\begin{array}{ccc|c} 25.5 & 0 & 0 & \\ 0 & 1.29 & 0 & \\ 0 & 0 & 0 & \\ 0 & 0 & 0 & \end{array} \right) \left(\begin{array}{ccc} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{array} \right)$$

- The first singular value is much bigger than the second, and so represents the vast majority of what A does (note, the vectors in U and V are unit length)
- Thus, one could approximate A quite well by only using the terms associated with the largest singular value
- This is not a valid factorization, but an approximation (and the idea behind PCA)

Summary

- The columns of V that do not correspond to “nonzero” singular values form an orthonormal basis for the null space of A
- The remaining columns of V form an orthonormal basis for the space perpendicular to the null space of A
- The columns of U corresponding to “nonzero” singular values form an orthonormal basis for the range of A
- The remaining columns of U form an orthonormal basis for the space perpendicular to the range of A
- One can drop the columns of U and V that do not correspond to “nonzero” singular values and still obtain a valid factorization of A
- One can drop the columns of U and V that correspond to “small/smaller” singular values and still obtain a reasonable approximation of A

Example (Wide Matrix)

$$A = \begin{pmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{pmatrix} =$$
$$\begin{pmatrix} .504 & -.761 & .408 \\ .574 & -.057 & -.816 \\ .644 & .646 & .408 \end{pmatrix} \begin{pmatrix} 25.5 & 0 & 0 & 0 \\ 0 & 1.29 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} .141 & .344 & .547 & .750 \\ .825 & .426 & .028 & -.371 \\ -.420 & .298 & .644 & -.542 \\ -.351 & .782 & -.509 & .079 \end{pmatrix}$$

- A maps from R^4 to R^3 and so has at least a 1D null space (green)
- The 3rd singular value is 0, and the associated (pink) terms make no contribution to the final result

Example (Wide Matrix)

$$A = \begin{pmatrix} 1 & 4 & 7 & 10 \\ 2 & 5 & 8 & 11 \\ 3 & 6 & 9 & 12 \end{pmatrix} =$$

$$\left(\begin{array}{ccc|c} .504 & -.761 & .408 & .408 \\ .574 & -.057 & -.816 & -.816 \\ .644 & .646 & .408 & .408 \end{array} \right) \left(\begin{array}{ccccc} 25.5 & 0 & 0 & 0 \\ 0 & 1.29 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} \right) \left(\begin{array}{cccc} .141 & .344 & .547 & .750 \\ .825 & .426 & .028 & -.371 \\ -.420 & .298 & .644 & .542 \\ -.351 & .782 & -.509 & .079 \end{array} \right)$$

- Only a 2D subspace of R^4 matters, with the rest of R^4 in the null space of A
- Only a 2D subspace of R^3 is in the range of A

Notes

- The SVD is often unwieldy for computational purposes
- However, replacing matrices by their SVD can be quite useful/enlightening for theoretical pursuits
- Moreover, its theoretical underpinnings are often used to devise computational algorithms
- The SVD is unique under certain assumptions, such as all $\sigma_k \geq 0$ and in descending order
- However, one can make both a σ_k and its associated column in U negative for an “alternate SVD” (see e.g. “Invertible Finite Elements For Robust Simulation of Large Deformation”, Irving et al. 2004)

Solving Linear Systems

- $Ac = b$ becomes $U\Sigma V^T c = b$ or $\Sigma(V^T c) = (U^T b)$ or $\Sigma\hat{c} = \hat{b}$
- The unknowns c are remapped into the space spanned by V , and the right hand side b is remapped into the space spanned by U
- **Every matrix is a diagonal matrix, when viewed in the right space**
- Solve the diagonal system $\Sigma\hat{c} = \hat{b}$ by dividing the entries of \hat{b} by the singular values σ_k ; then, $c = V\hat{c}$
- The SVD transforms the problem into an inherently diagonal space with eigenvectors along the coordinate axes
- Circles becoming ellipses (discussed earlier) is still problematic
 - Eccentricity is caused by ratios of singular values (since U and V are orthogonal matrices)

Condition Number

- The condition number of A is $\frac{\sigma_{max}}{\sigma_{min}}$ and measures closeness to being singular
- For a square matrix, it measures the difficulty in solving $Ac = b$
- For a rectangular (and square) matrix, it measures how close the columns are to being linearly independent
- The condition number does not depend on the right hand side
- The condition number is always bigger than 1, and approaches ∞ for nearly singular matrices
- Singular matrices have condition number equal to ∞ , since $\sigma_{min} = 0$

Singular Matrices

- Diagonalize $Ac = b$ to $\Sigma(V^T c) = (U^T b)$, e.g. $\begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \end{pmatrix}$ with $\hat{c}_1 = \frac{\hat{b}_1}{\sigma_1}$, $\hat{c}_2 = \frac{\hat{b}_2}{\sigma_2}$
- When $\sigma_2 = 0$, there is no unique solution:
 - When $\hat{b}_2 = 0$, there are infinite solutions for \hat{c}_2 (but \hat{c}_1 is still uniquely determined)
 - When $\hat{b}_2 \neq 0$, there is no solution for \hat{c}_2 , and b is not in the range of A (but \hat{c}_1 is still uniquely determined)
- Consider: $\begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \\ \hat{c}_3 \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{pmatrix}$ which still has $\hat{c}_1 = \frac{\hat{b}_1}{\sigma_1}$, $\hat{c}_2 = \frac{\hat{b}_2}{\sigma_2}$
 - When $\hat{b}_3 = 0$, the last row adds no new information
 - When $\hat{b}_3 \neq 0$, the last row is false and there is no solution
- Consider: $\begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \sigma_2 & 0 \end{pmatrix} \begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \\ \hat{c}_3 \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{pmatrix}$ which still has $\hat{c}_1 = \frac{\hat{b}_1}{\sigma_1}$, $\hat{c}_2 = \frac{\hat{b}_2}{\sigma_2}$
 - Infinite solutions work for \hat{c}_3

Understanding Variables

- When $\sigma_k \neq 0$, one can state a value for \hat{c}_k
 - When $\sigma_k = 0$ or there is no σ_k , there is no data for \hat{c}_k
 - This does not mean that other parameters cannot be adequately determined
-
- Consider a row i of Σ that is identically zero
 - When $\hat{b}_i = 0$, this row indicates that there is extra redundant data
 - When $\hat{b}_i \neq 0$, this row indicates that there is conflicting information in the data
 - Conflicting information doesn't necessarily imply that all is lost, i.e. "no solution"; rather, it might merely mean that the data contains a bit of noise
 - Regardless, in spite of the conflicting information, the determinable \hat{c}_k represent the "best" that one can do

Norms

- Common norms: $\|c\|_1 = \sum_k |c_k|$, $\|c\|_2 = \sqrt{\sum_k c_k^2}$, $\|c\|_\infty = \max_k |c_k|$
- "All norms are interchangeable" is a theoretically valid statement (**only**)
- In practice, the "worst case scenario" (L^∞) and the "average" (L^1 , L^2 , etc.) are not interchangeable
 - E.g. $(100 \text{ people} * 98.6^\circ + 1 \text{ person} * 105^\circ) / (101 \text{ people}) = 98.66^\circ$
 - Their average temperature is 98.66° , so everyone is fine?

Matrix Norms

- Define the norm of a matrix $\|A\| = \max_{c \neq 0} \frac{\|Ac\|}{\|c\|}$, so:
 - $\|A\|_1$ is the maximum absolute value column sum
 - $\|A\|_\infty$ is the maximum absolute value row sum
 - $\|A\|_2$ is the square root of the maximum eigenvalue of $A^T A$, i.e. the maximum singular value of A
- The condition number for solving $Ac = b$ is $\|A\|_2 \|A^{-1}\|_2$
- Since $A^{-1} = V\Sigma^{-1}U^T$ where Σ^{-1} has diagonal entries $\frac{1}{\sigma_k}$, $\|A^{-1}\|_2 = \frac{1}{\sigma_{min}}$
- Thus, $\|A\|_2 \|A^{-1}\|_2 = \frac{\sigma_{max}}{\sigma_{min}}$

Unit 4

Special Matrices

(Strict) Diagonal Dominance

- The magnitude of each diagonal element is (either):
 - strictly larger than the sum of the magnitudes of all the other elements in its row
 - strictly larger than the sum of the magnitudes of all the other elements in its column
- One may row/column scale and permute rows/columns to achieve diagonally dominance (since it is just a rewriting of the equations)
 - Recall: **choosing the form of the equations wisely is important**
- E.g. consider $\begin{pmatrix} 3 & -2 \\ 5 & 1 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 9 \\ 4 \end{pmatrix}$
- Switch rows $\begin{pmatrix} 5 & 1 \\ 3 & -2 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 9 \end{pmatrix}$ and column scale $\begin{pmatrix} 5 & -2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} c_1 \\ -.5c_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 9 \end{pmatrix}$

(Strict) Diagonal Dominance

- Strictly diagonally dominant (square) matrices are guaranteed to be non-singular
- Since $\det(A) = \det(A^T)$, either row or column diagonal dominance is enough
- Column diagonal dominance guarantees pivoting is not required during *LU* factorization
- However, pivoting still improves robustness
- E.g. consider $\begin{pmatrix} 4 & 3 \\ -2 & 50 \end{pmatrix}$ where 50 is more desirable than 4 for a_{11}

Inner Product

- Consider the space of all vectors with length m
- The dot/inner product of two vectors is $u \cdot v = \sum_i u_i v_i$
- The magnitude of a vector is $\|v\|_2 = \sqrt{v \cdot v} (\geq 0)$
- Alternative notations: $\langle u, v \rangle = u \cdot v = u^T v$
- Weighted inner product defined via an $n \times n$ matrix A
- $\langle u, v \rangle_A = u \cdot A v = u^T A v$
- Since $\langle v, u \rangle_A = v^T A u = u^T A^T v$, weighted inner products commute when A is symmetric
- The standard dot product uses identity matrix weighting: $\langle u, v \rangle = \langle u, v \rangle_{I^m}$

Definiteness

- Assume A is symmetric so that $\langle u, v \rangle_A = \langle v, u \rangle_A$
- A is positive definite if and only if $\langle v, v \rangle_A = v^T A v > 0$ for $\forall v \neq 0$
- A is positive semi-definite if and only if $\langle v, v \rangle_A = v^T A v \geq 0$ for $\forall v \neq 0$
- We abbreviate with SPD and SP(S)D
- A is negative definite if and only if $\langle v, v \rangle_A = v^T A v < 0$ for $\forall v \neq 0$
- A is negative semi-definite if and only if $\langle v, v \rangle_A = v^T A v \leq 0$ for $\forall v \neq 0$
- If A is negative (semi) definite, then $-A$ is positive (semi) definite (and vice versa)
- Thus, can convert such problems to SPD/SP(S)D instead
- A is considered indefinite when it is neither positive/negative semi-definite

Eigenvalues

- SPD matrices have all eigenvalues > 0
- SP(S)D matrices have all eigenvalues ≥ 0
- Symmetric negative definite matrices have all eigenvalues < 0
- Symmetric negative semi-definite matrices have all eigenvalues ≤ 0
- Indefinite matrices have both positive and negative eigenvalues

SVD Construction (Important Detail)

- Let $A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ so that $A^T A = AA^T = I$, and thus $U = V = \Sigma = I$
- But $A \neq U\Sigma V^T = I$ **What's wrong?**
- Given a column vector v_k of V , $Av_k = U\Sigma V^T v_k = U\Sigma \hat{e}_k = U\sigma_k \hat{e}_k = \sigma_k u_k$ where u_k is the corresponding column of U
- $Av_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} = u_1$ but $Av_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 1 \end{pmatrix} = u_2$
- Although eigenvectors may be scaled by an arbitrary constant, the constraint that U and V be orthonormal forces their columns to be unit length
- However, there are still two choices for the direction of each column
- Multiplying u_2 by -1 makes $U = A$ and thus $A = U\Sigma V^T$ as desired

SVD Construction (Important Detail)

- An orthogonal matrix has determinant equal to ± 1 , where -1 indicates a reflection of the coordinate system
- If $\det V = -1$, flip the direction of any column to make $\det V = 1$ (so V does not contain a reflection)
- Then, for each v_k , compare Av_k to $\sigma_k u_k$ and flip the direction of u_k when necessary in order to make $Av_k = \sigma_k u_k$
- $\det U = \pm 1$ and may contain a reflection
- When $\det U = -1$, one can flip the sign of the smallest singular value in Σ to be negative, whilst also flipping the direction of the corresponding column in U so that $\det U = 1$
- This embeds the reflection into Σ and is called the polar-SVD (Irving et al. 2004)

Symmetric Matrices (SVD)

- Since $A^T A = AA^T = A^2$, both the columns of U and the columns of V are eigenvectors of A^2
- They are identical (but potentially opposite) directions: $u_k = \pm v_k$
- Thus, $A v_k = \sigma_k u_k$ implies $A v_k = \pm \sigma_k v_k$
- That is, the v_k (and u_k) are eigenvectors of A with eigenvalues $\pm \sigma_k$
- Similar to the polar SVD, pull negative signs out of the columns of U into the σ_k to obtain $U = V$ and $A = V \Lambda V^T$ as a modified SVD
- $A = V \Lambda V^T$ implies $AV = V \Lambda$ which is the matrix form of the eigensystem of A
- Thus, Λ contains the positive and negative eigenvalues of A

SPD Matrices

- When A is SP(S)D, $\Lambda = \Sigma$ and the standard SVD is $A = V\Sigma V^T$ (i.e. $U = V$)
- The singular values are the (all positive) eigenvalues of A (since $AV = V\Sigma$)
- Constructing V with $\det V = 1$, all $\sigma_k > 0$ implies that there are no reflections
- Since all $\sigma_k > 0$, the matrix has full rank and is invertible
- SP(S)D (and not SPD) has at least one $\sigma_k = 0$ and a null space
- Often, one can use modified SPD techniques for SP(S)D matrices
- Unfortunately, indefinite matrices are significantly more challenging

Making/Breaking Symmetry

- Row/column scaling breaks symmetry:
 - Row scaling $\begin{pmatrix} 5 & 3 \\ 3 & -4 \end{pmatrix}$ by -2 gives a non-symmetric $\begin{pmatrix} 5 & 3 \\ -6 & 8 \end{pmatrix}$
 - Additional column scaling by -2 gives $\begin{pmatrix} 5 & -6 \\ -6 & -16 \end{pmatrix}$
- Scaling the same row/column together in the same way preserves symmetry
- Important: a nonsymmetric matrix might be inherently symmetric when properly rescaled/rearranged

Rules Galore

- There are many rules/theorems regarding special matrices (especially for SPD)
- Important to be aware of reference material (and to look things up)
- Examples:
 - SPD matrices don't require pivoting during LU factorization
 - A symmetric (strictly) diagonally dominant matrix with positive diagonal entries is positive definite
 - Jacobi and Gauss-Seidel iteration converge when a matrix is strictly (or irreducibly) diagonally dominant
 - Etc.

Cholesky Factorization

- SPD matrices have LU factorization of LL^T and don't require elimination to find it

- Consider $\begin{pmatrix} a_{11} & a_{21} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} \\ 0 & l_{22} \end{pmatrix} = \begin{pmatrix} l_{11}^2 & l_{11}l_{21} \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 \end{pmatrix}$

- Then $l_{11} = \sqrt{a_{11}}$ and $l_{21} = \frac{a_{21}}{l_{11}}$ and $l_{22} = \sqrt{a_{22} - l_{21}^2}$

`for(j=1,n){
 for(k=1,j-1) for(i=j,n) aij -= aikajk;
 ajj = sqrt(ajj); for(k=j+1,n) akj/= ajj;`}

\\ For each column j of the matrix

\\ Loop over all previous columns k, and subtract a multiple of column k from the current column j

\\ Take the square root of the diagonal entry, and scale column j by that value

- This algorithm factors the matrix “in place” replacing A with L

Incomplete Cholesky Preconditioner

- Cholesky factorization can be used to construct a preconditioner for a sparse matrix
- The full Cholesky factorization would fill in too many non-zero entries
- So, incomplete Cholesky preconditioning uses Cholesky factorization with the caveat that only the nonzero entries are modified (all zeros remain zeros)

Symmetric Approximation

- For non-symmetric A , a symmetric $\hat{A} = \frac{1}{2}(A + A^T)$ averages off-diagonal components
- Solving the symmetric $\hat{A}c = b$ instead of the non-symmetric $Ac = b$ gives a faster/easier (though error prone) approximation to a problem that might not require too much accuracy
- Alternatively, the inverse of \hat{A} (or the notion thereof) may be used to devise a preconditioner for $Ac = b$

Unit 5

Iterative Solvers

Iterative vs. Direct Solvers

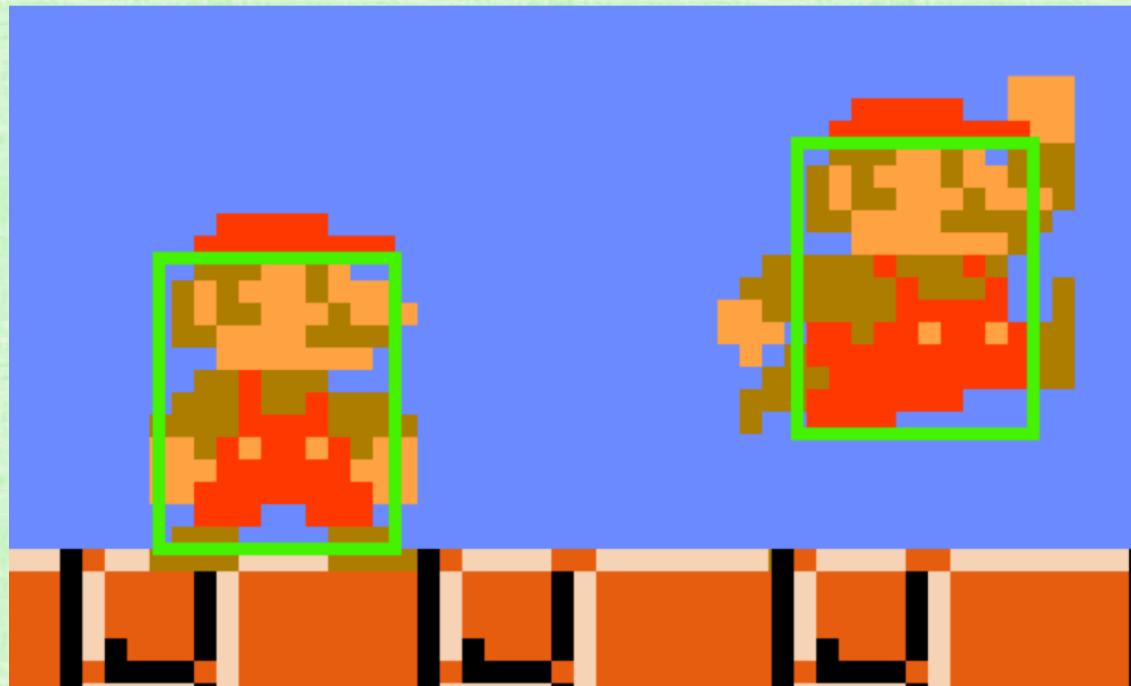
- Direct Solver/Method – closed form strategy, e.g. quadratic/Cardano formula, Gaussian Elimination for LU factorization, Cholesky factorization, etc.
- Iterative Solver/Method
 - start with initial guess c^1
 - use recursive approach to improve that guess: c^2, c^3, c^4, \dots
 - terminate based on a stopping criterion, e.g. when error is small $\|c^q - c^{exact}\| \leq \epsilon$
- A direct method can be used to obtain an initial guess
- Iterative methods are great for sparse matrices, as they often can ignore 0 entries
 - E.g. by formulating the method via the matrix's action on a vector
- Direct solvers are more commonly used on dense matrices
- **Iterative solvers are used for training Neural Networks!**

Issues with Direct Methods

- (Recall) Quadratic formula loses precision, and can fail, when $-b \pm \sqrt{b^2 - 4ac}$ has catastrophic cancellation
 - The de-rationalized quadratic formula instead uses $-b \mp \sqrt{b^2 - 4ac}$
 - Using one formula for each root avoids catastrophic cancellation
- Cardano's formula for the roots of cubic equations suffers from similar issues, but there is no good way to fix the formula
- The computed roots/solutions too often have unacceptably high error
- To demonstrate the need for more accurate cubic roots, let's consider collision detection

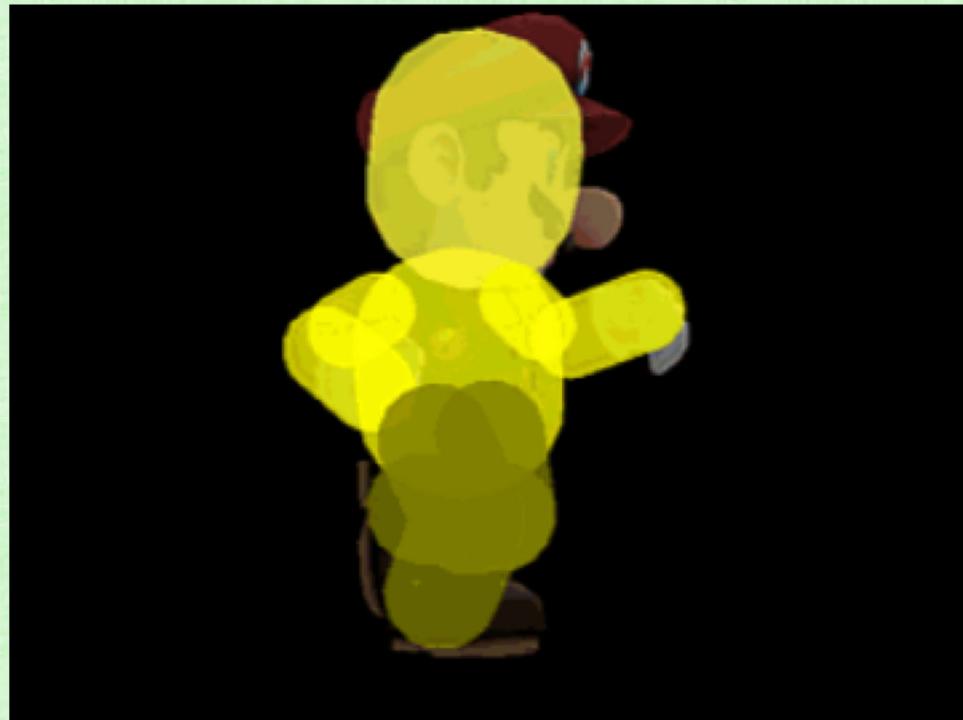
Hit Box

- In order to detect interactions between objects in video games, objects were assigned a hit box
- Anything inside an object's hit box can potentially interact with (i.e. hit) it



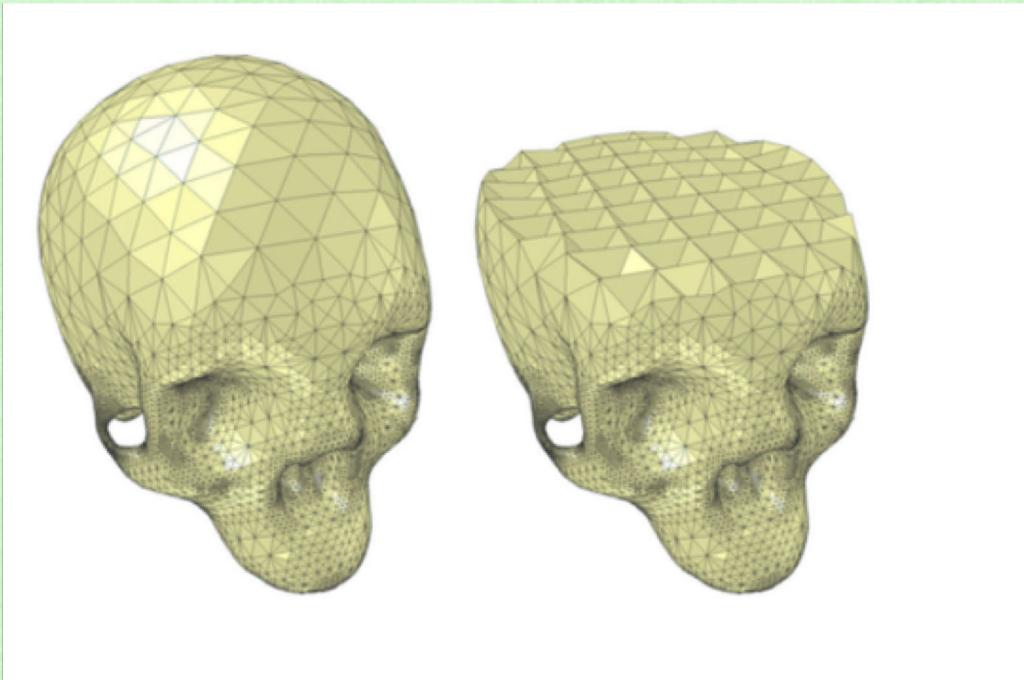
Better Hit Boxes

- These evolved over time to more complicated shapes in both 2D and 3D
 - e.g. spheres, ellipsoids, capsules, etc.
- Anything inside any of an object's hit boxes can potentially interact with it



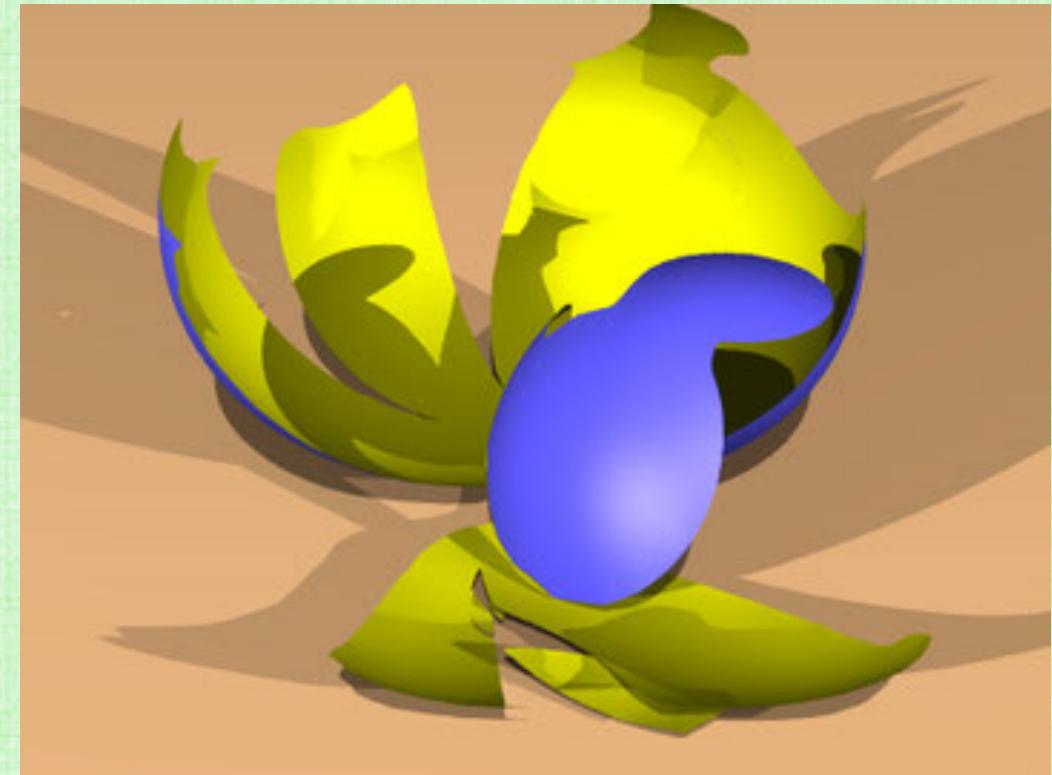
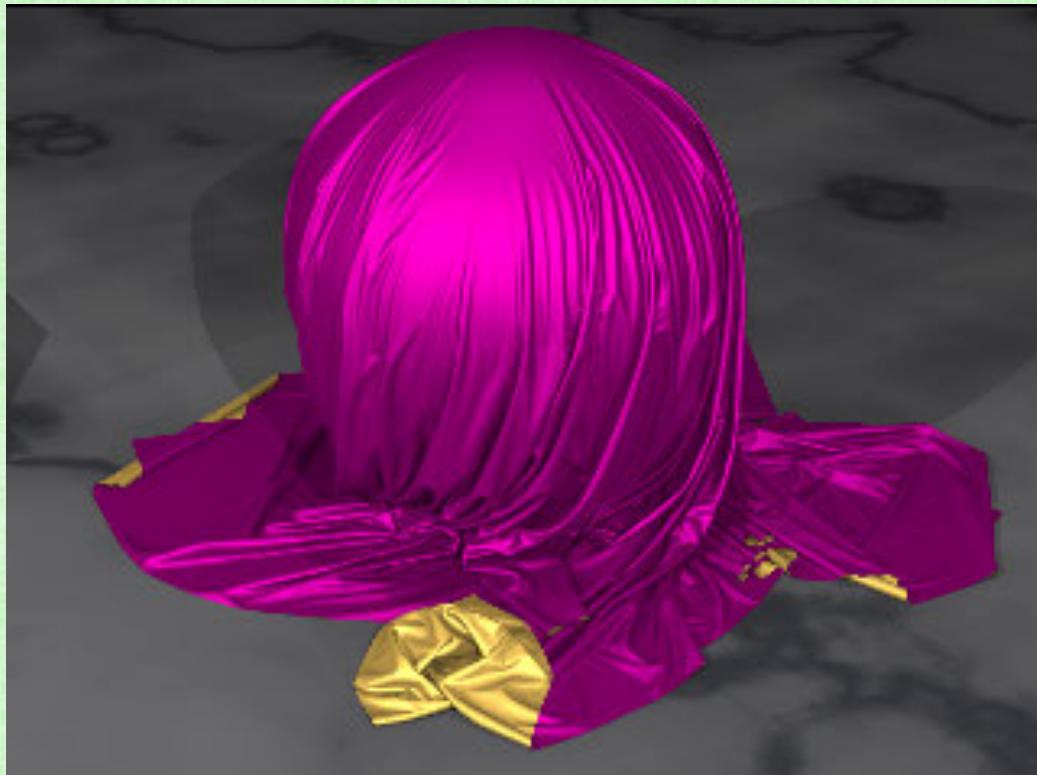
Accurate Collision Detection

- More complex objects are often modeled by triangle surface meshes
- The interior can be filled with tetrahedra, or approximated with other objects
- Anything inside any of an object's interior structures can potentially interact with it



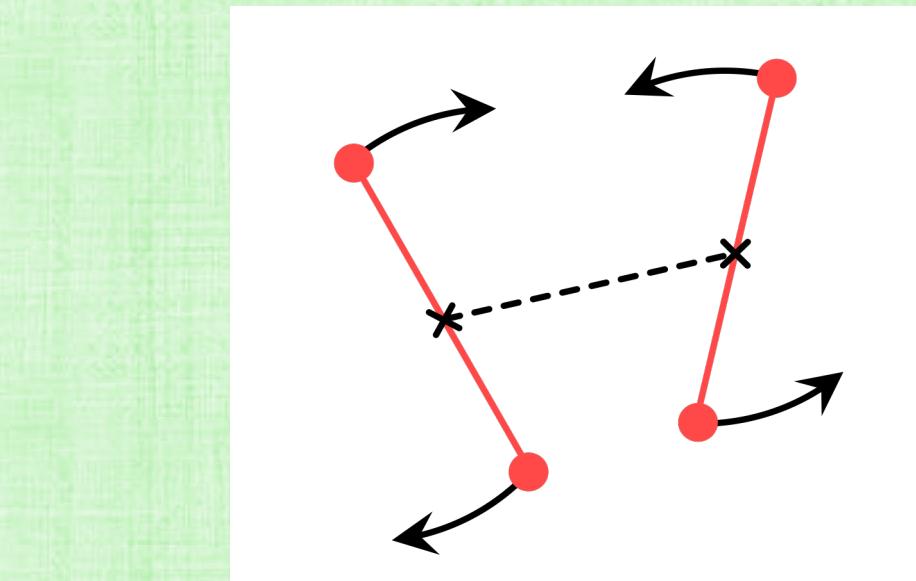
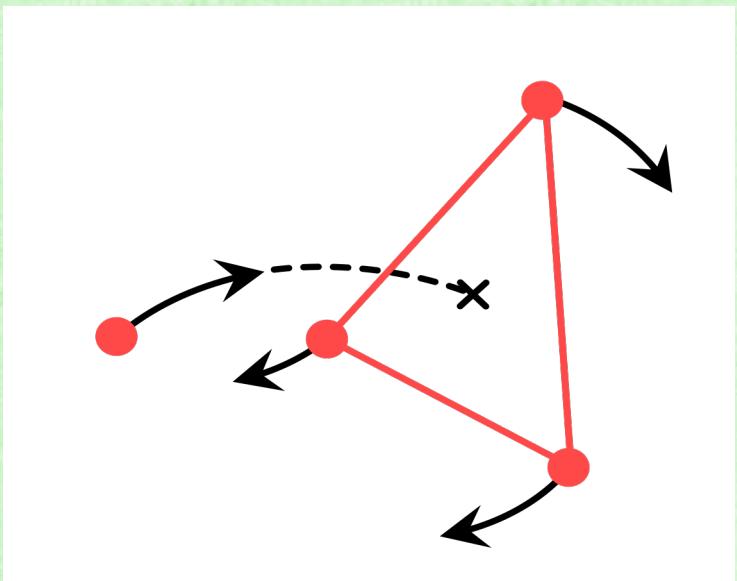
Objects Without Interiors

- Very thin objects, such as cloth/shells, do not have an interior region
- One cannot use the same concept of inside to detect potential interactions



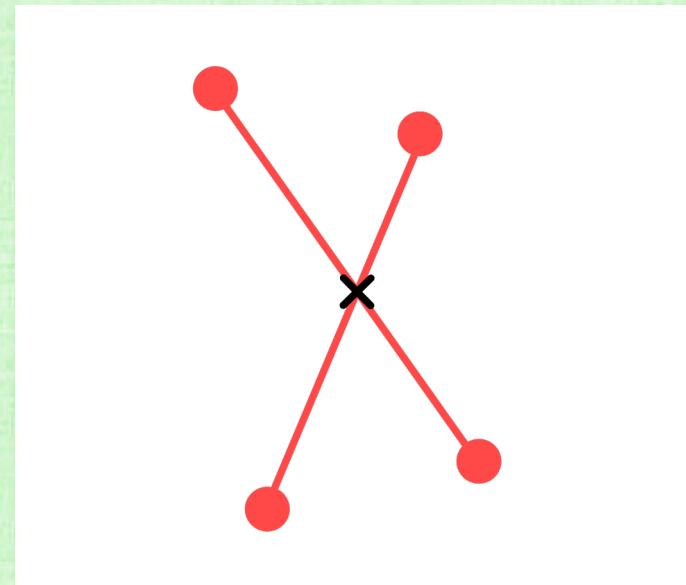
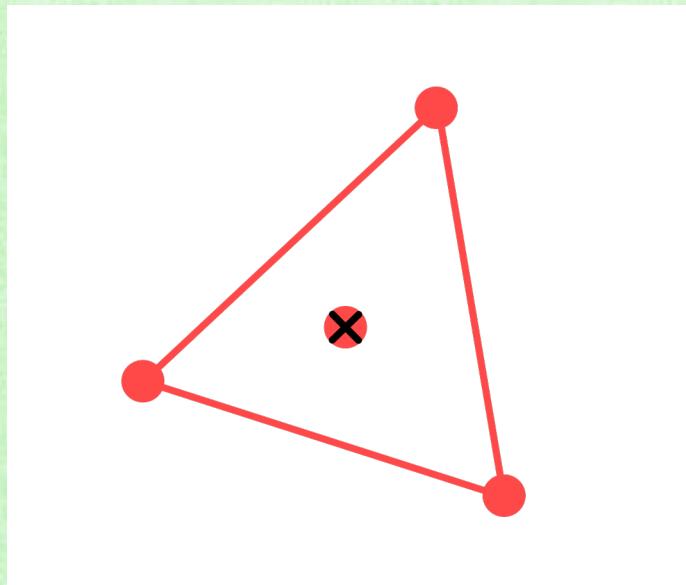
Continuous Collision Detection (CCD)

- Model time varying trajectories of surface triangle vertices (or other geometry) to see if/when they collide with each other
- Doesn't depend on an interior region
- Two cases to consider for triangles: (1) Point-Face, (2) Edge-Edge



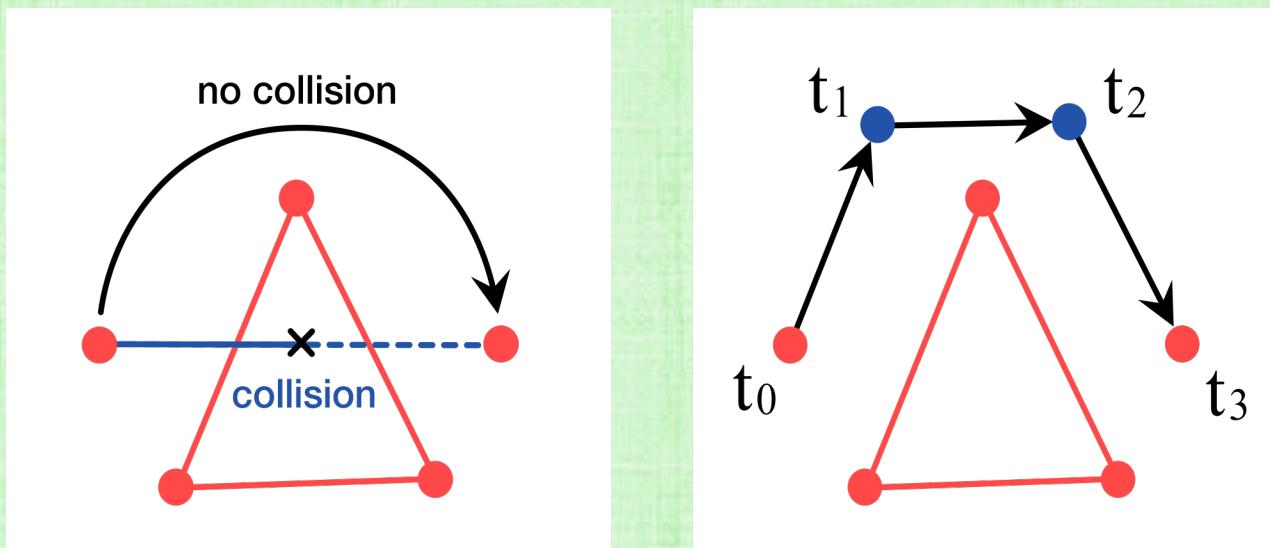
Continuous Collision Detection (CCD)

- Both the Point-Face case and the Edge-Edge case require their 4 relevant points to become coplanar in order to potentially/actually collide
- Once deemed coplanar, a second check determines whether the lone point is inside the triangle (for Point-Face) or the two edges intersect (for Edge-Edge)



Continuous Collision Detection (CCD)

- Consider time t_o to time t_f and assume all the points have constant velocities $V_i(t_o)$ for $i = 1, 2, 3, 4$
- Then, their time evolving positions are: $X_i(t) = X_i(t_o) + V_i(t_o)t$ for $t_o \leq t \leq t_f$
- Although their paths are (generally) curved, considering piecewise linear increments is sufficient for preventing self-intersecting states



Continuous Collision Detection (CCD)

- The equation/condition for the positions $X_i(t)$ of all four points to be co-planar is a cubic equation in t
- Need to find the roots of this cubic equation in the interval $[t_o, t_f]$
- Cubic equation solvers are so error prone that collisions are often missed, and the cloth/shell ends up in a spurious self-intersecting state
- A very carefully devised/implemented iterative solver for cubic equations was able to detect all collisions
 - It requires double precision (and fails too often in single precision)
 - See Bridson et al. “Robust Treatment of Collisions, Contact, and Friction for Cloth Animation” (2002)

Residual (and Error)

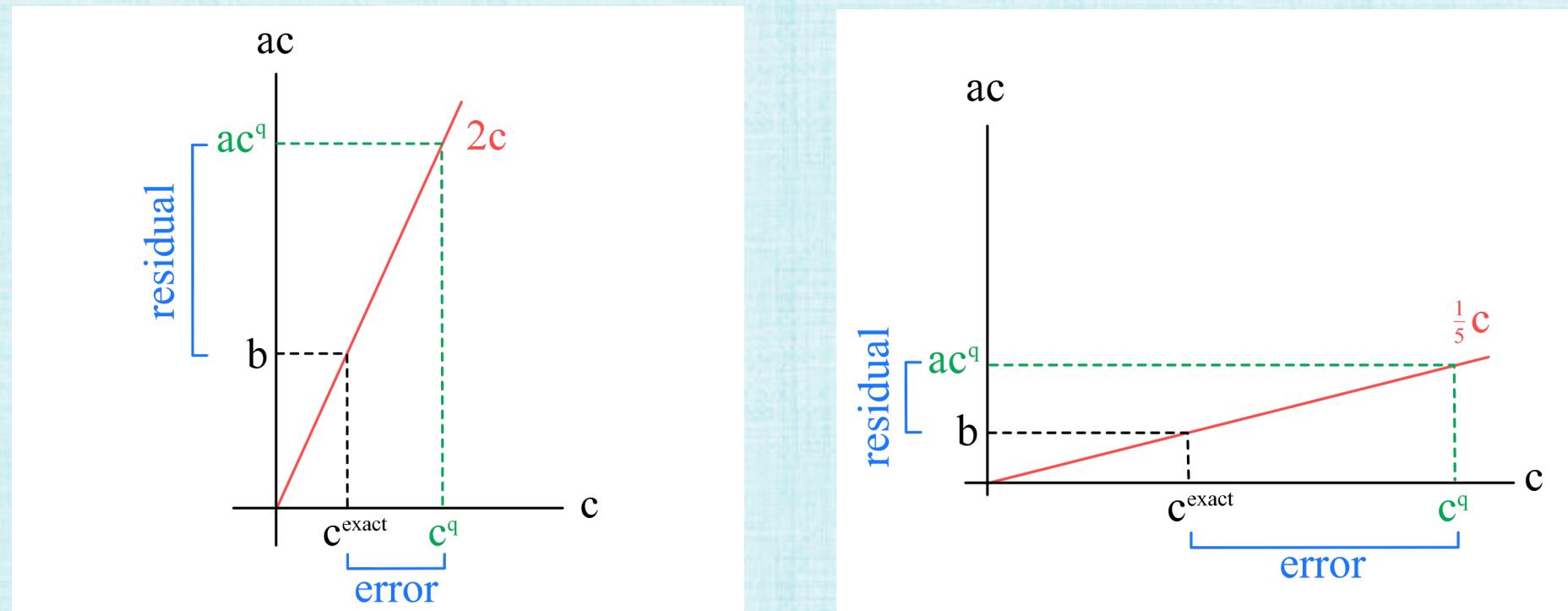
- When solving $Ac = b$, given a guess c^q , the residual is $r^q = b - Ac^q$
- The residual measures the errors in the equations, not the error in the solution
- Given error $e^q = c^q - c^{exact}$, one has:

$$r^q = b - Ac^q = Ac^{exact} - Ac^q = A(c^{exact} - c^q) = -Ae^q$$

- That is, the residual is the error transformed into the space that b lives in (the range of A)

Residual (1D example)

- Consider a simple size 1×1 matrix, i.e. $[a]c = b$ with exact solution $c = \frac{b}{a}$
- Since $r^q = -ae^q$, smaller a values lead to deceptively small residuals even when the error is large



Residual

- "All matrices are diagonal matrices", and diagonal matrices represent decoupled 1D scalar problems
- Using the SVD, $r^q = -Ae^q$ becomes $(U^T r^q) = -\Sigma(V^T e^q)$ which is a decoupled set of diagonal equations
- Each decoupled equation has the form $\hat{r}_k^q = -\sigma_k \hat{e}_k^q$ (seen on the previous slide)
- Small σ_k lead to deceptively small residuals even when the error is large
- A small residual indicates a small error for larger singular values, but not for smaller singular values

Line Search (in parameter space)

- Choose a search direction s^q and move some distance α^q in that direction to find the next iterative guess: $c^{q+1} = c^q + \alpha^q s^q$
 - There are various strategies for choosing α^q including the notion of safe sets that clamp its maximum attainable value
 - Subtracting c^{exact} from both sides of this recursion gives $e^{q+1} = e^q + \alpha^q s^q$, and multiplying through by A gives $r^{q+1} = r^q - \alpha^q A s^q$
- Preferably, one would follow s^q until no error was left in that direction, i.e. until the remaining error was orthogonal to s^q , i.e. $e^{q+1} \cdot s^q = 0$
- Since the error is unknown (otherwise the solution is known), one can instead progress until the residual is orthogonal to s^q , i.e. $r^{q+1} \cdot s^q = 0$
 - Plugging in the recursion for r^{q+1} gives $\alpha^q = \frac{s^q \cdot r^q}{s^q \cdot A s^q}$

Steepest Descent

- Steepest descent chooses the search direction to be the steepest downhill direction, which turns out (in this case) to be the residual, i.e. $s^q = r^q$
- So $r^q = b - Ac^q$, $\alpha^q = \frac{r^q \cdot r^q}{r^q \cdot Ar^q}$, $c^{q+1} = c^q + \alpha^q r^q$ is iterated until r^q is considered small enough
- Note that $r^q = b - Ac^q$ can be replaced with $r^q = r^{q-1} - \alpha^{q-1} Ar^{q-1}$ where Ar^{q-1} had already been computed to find α^{q-1} (this eliminates a possibly expensive multiplication by A)
- The main drawback to steepest descent is that it repeatedly searches in the same directions too often, especially for higher condition number matrices (much more on this later)

Conjugate Gradients (CG) Method

- A very efficient/robust method for SPD systems
- Converges (theoretically) in n -steps for an $n \times n$ matrix
 - Actually, converges in the number of steps equal to the number of distinct eigenvalues
 - Almost converges in the number of steps equal to the number of eigenvalue clusters
 - Thus, preconditioning makes a big difference, if it can cluster eigenvalues
- Motivation: choosing ***orthogonal*** search directions would preclude repeatedly searching in the same directions as Steepest Descent ineffectually does
- Turns out to be difficult to implement this orthogonality
- Instead: choose search directions to be A-orthogonal
- That is, $\langle s^q, s^{\hat{q}} \rangle_A = 0$ for $q \neq \hat{q}$, instead of $\langle s^q, s^{\hat{q}} \rangle = 0$

Gram-Schmidt

- Orthogonalizes a set of vectors
- For each vector, subtract its dot product overlap with all prior vectors, making it orthogonal to them
- A-orthogonal Gram-Schmidt simply uses A-weighted dot products
- Given vector \bar{S}^q , subtract out the A-overlap with s^1 to s^{q-1} so that the resulting vector s^q has $\langle s^q, s^{\hat{q}} \rangle_A = 0$ for $\hat{q} \in \{1, 2, \dots, q-1\}$
- That is, $s^q = \bar{S}^q - \sum_{\hat{q}=1}^{q-1} \frac{\langle \bar{S}^q, s^{\hat{q}} \rangle_A}{\langle s^{\hat{q}}, s^{\hat{q}} \rangle_A} s^{\hat{q}}$ where the two non-normalized $s^{\hat{q}}$ both require division by their norm (note that $\langle s^{\hat{q}}, s^{\hat{q}} \rangle_A = \|s^{\hat{q}}\|_A^2$)
- Proof: $\langle s^q, s^{\hat{q}} \rangle_A = \langle \bar{S}^q, s^{\hat{q}} \rangle_A - \frac{\langle \bar{S}^q, s^{\hat{q}} \rangle_A}{\langle s^{\hat{q}}, s^{\hat{q}} \rangle_A} \langle s^{\hat{q}}, s^{\hat{q}} \rangle_A = 0$

CG (Error Analysis)

- In the A-orthogonal basis of search directions, the initial error is $e^1 = \sum_{\hat{q}} \beta^{\hat{q}} s^{\hat{q}}$ so $\langle s^q, e^1 \rangle_A = \beta^q \langle s^q, s^q \rangle_A$
- Error recursion gives $e^q = e^1 + \sum_{\hat{q}=1}^{q-1} \alpha^{\hat{q}} s^{\hat{q}}$, so $\langle s^q, e^q \rangle_A = \langle s^q, e^1 \rangle_A$ since $\langle s^q, s^{\hat{q}} \rangle_A = 0$ for $q \neq \hat{q}$
- Recall that progressing until $r^{q+1} \cdot s^q = 0$ gave $\alpha^q = \frac{s^q \cdot r^q}{s^q \cdot A s^q}$ ($= -\frac{\langle s^q, e^q \rangle_A}{\langle s^q, s^q \rangle_A}$)
- Thus $\alpha^q = -\beta^q$, so that $e^1 = \sum_{\hat{q}} (-\alpha^{\hat{q}}) s^{\hat{q}}$ and $e^q = \sum_{\hat{q}=q}^n (-\alpha^{\hat{q}}) s^{\hat{q}}$, proving that the error is indeed cancelled out in n steps!
- Aside: For $\tilde{q} < q$, one has $s^{\tilde{q}} \cdot r^q = -\langle s^{\tilde{q}}, e^q \rangle_A = 0$ implying that **the residual is orthogonal to all previous search directions** (not just the previous one)

CG (Gram-Schmidt)

- Choose candidate search directions $\bar{S}^q = r^q$, and make A-orthogonal via Gram-Schmidt
- That is, $s^q = r^q - \sum_{\hat{q}=1}^{q-1} \frac{\langle r^q, s^{\hat{q}} \rangle_A}{\langle s^{\hat{q}}, s^{\hat{q}} \rangle_A} s^{\hat{q}}$
- Dot product this with $r^{\tilde{q}}$ to get $s^q \cdot r^{\tilde{q}} = r^q \cdot r^{\tilde{q}} - \sum_{\hat{q}=1}^{q-1} \frac{\langle r^q, s^{\hat{q}} \rangle_A}{\langle s^{\hat{q}}, s^{\hat{q}} \rangle_A} s^{\hat{q}} \cdot r^{\tilde{q}}$
 - When $\tilde{q} > q$, one has $0 = r^q \cdot r^{\tilde{q}} + 0$ implying that the residual are all orthogonal
 - When $\tilde{q} = q$, one has $s^q \cdot r^q = r^q \cdot r^q + 0$, so that $\alpha^q = \frac{r^q \cdot r^q}{\langle s^q, s^q \rangle_A}$
- Starting with the recursion $r^{q+1} = r^q - \alpha^q A s^q$
- Dot product this with $r^{\tilde{q}}$ to get $r^{\tilde{q}} \cdot r^{q+1} = r^{\tilde{q}} \cdot r^q - \alpha^q \langle r^{\tilde{q}}, s^q \rangle_A$
 - When $\tilde{q} = q + 1$, we get $r^{q+1} \cdot r^{q+1} = 0 - \alpha^q \langle r^{q+1}, s^q \rangle_A$ for the last term in the sum
 - When $\tilde{q} > q + 1$, we get $0 = 0 - \alpha^q \langle r^{\tilde{q}}, s^q \rangle_A$, so only the last term in the sum is nonzero
- Finally, $s^q = r^q + \frac{r^q \cdot r^q}{\alpha^{q-1} \langle s^{q-1}, s^{q-1} \rangle_A} s^{q-1} = r^q + \frac{r^q \cdot r^q}{r^{q-1} \cdot r^{q-1}} s^{q-1}$

CG (Method)

- Start with: $s^1 = r^1 = b - Ac^1$
- Iterate:
 - $\alpha^q = \frac{r^q \cdot r^q}{\langle s^q, s^q \rangle_A}$
 - $c^{q+1} = c^q + \alpha^q s^q$ and $r^{q+1} = r^q - \alpha^q As^q$ (both as usual)
 - $s^{q+1} = r^{q+1} + \frac{r^{q+1} \cdot r^{q+1}}{r^q \cdot r^q} s^q$
- Note: Because of Gram-Schmidt drift, one the search directions can become less A-orthogonal/effective over time; thus, occasionally throw out all search directions and start over with $s^1 = r^1 = b - Ac^1$

Non-Symmetric and Indefinite

- GMRES, MINRES, BiCGSTAB, etc...
- Generally speaking, iterative methods for non-symmetric and/or indefinite matrices are less stable, more error prone, and slower than CG

Unit 6

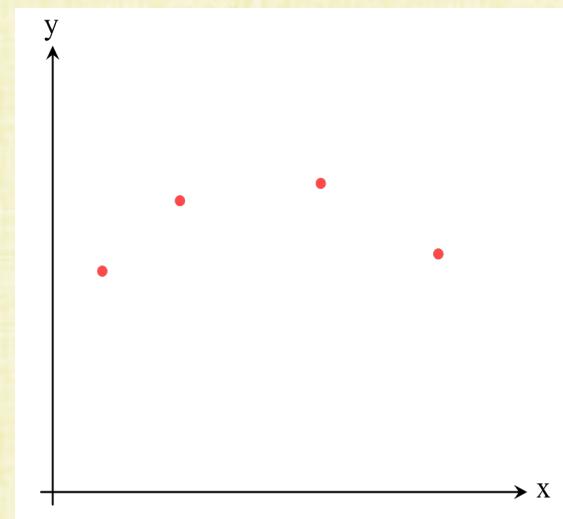
Local Approximations

Taylor Expansion

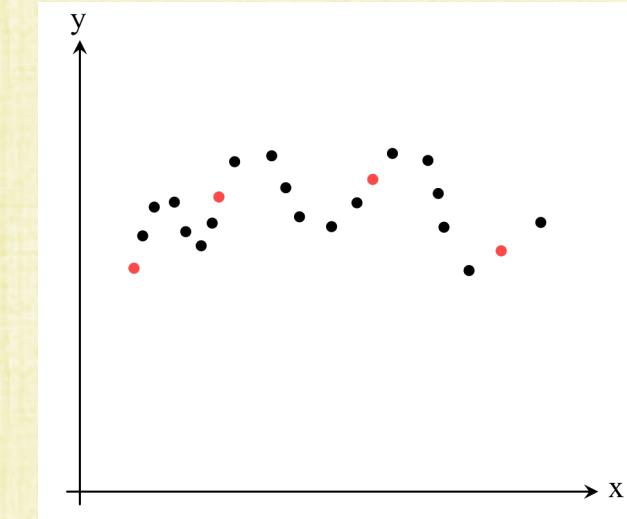
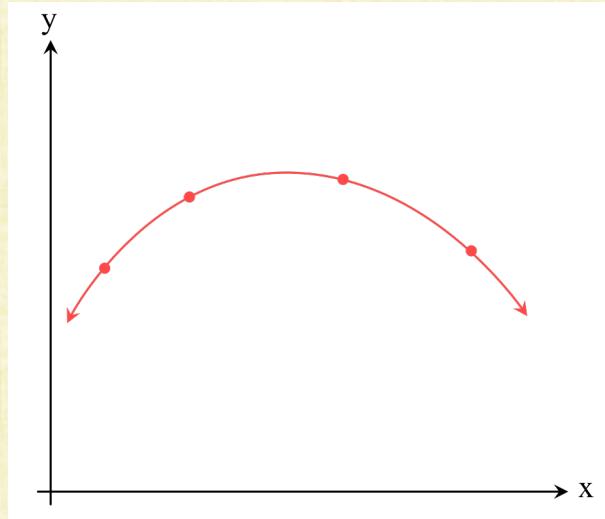
- $f(x + h) = \sum_{p=0}^{\infty} \frac{h^p}{p!} f^{(p)}(x) = \sum_{p=0}^{\hat{p}} \frac{h^p}{p!} f^{(p)}(x) + O(h^{\hat{p}+1})$
- Since the derivatives are bounded, $O(h^{\hat{p}+1}) \rightarrow 0$ as $h \rightarrow 0$
- Examples:
 - $f(x + h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + O(h^4)$ forward difference
 - $f(x - h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(x) + O(h^4)$ backward difference
- Approximations (truncated Taylor expansions) become more valid as $h \rightarrow 0$
 - $f(x + h) \approx f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x)$
 - $f(x - h) \approx f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(x)$

Sampling

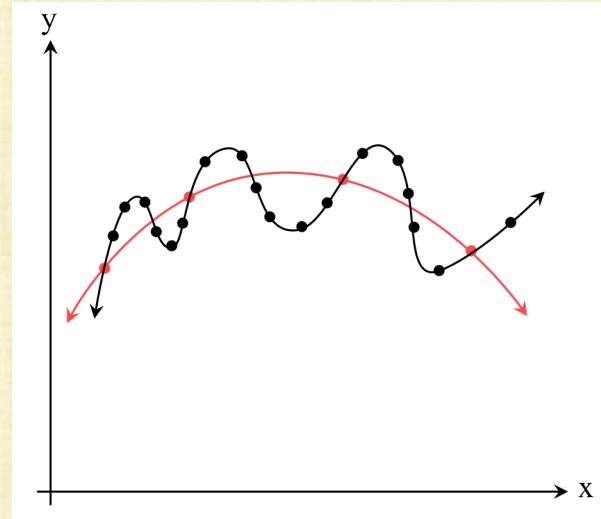
- Accurate approximate of a function is often limited by the amount of available data
- Given too few samples (left) one may "hallucinate" an incorrect function
- Adding more data allows better/proper feature resolution (right)
- Given "enough" sample points, a function tends to not vary too much in between them



under-resolved



resolved better with more data

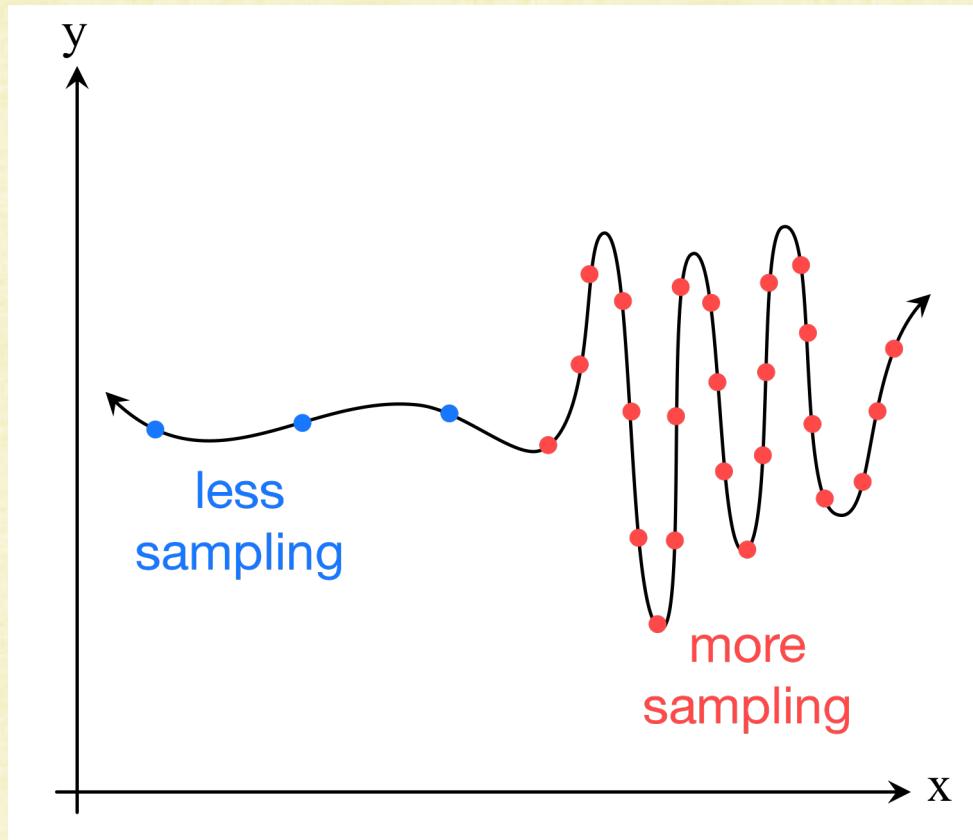


Well-Resolved Functions

- The Taylor expansion approximates a function f at a new location $x + h$ based on known information at a nearby point x
- When the sample points are “closely” spaced, the new location is “close” to a known sample point making h “small”
- However, large derivative values can overwhelm a small h
- Thus, functions with more variation need higher sampling rates
- (Also) Thus, smoother functions may utilize lower sampling rates
- Well-resolved functions have vanishing high order terms in their Taylor expansion making truncated Taylor expansions more valid

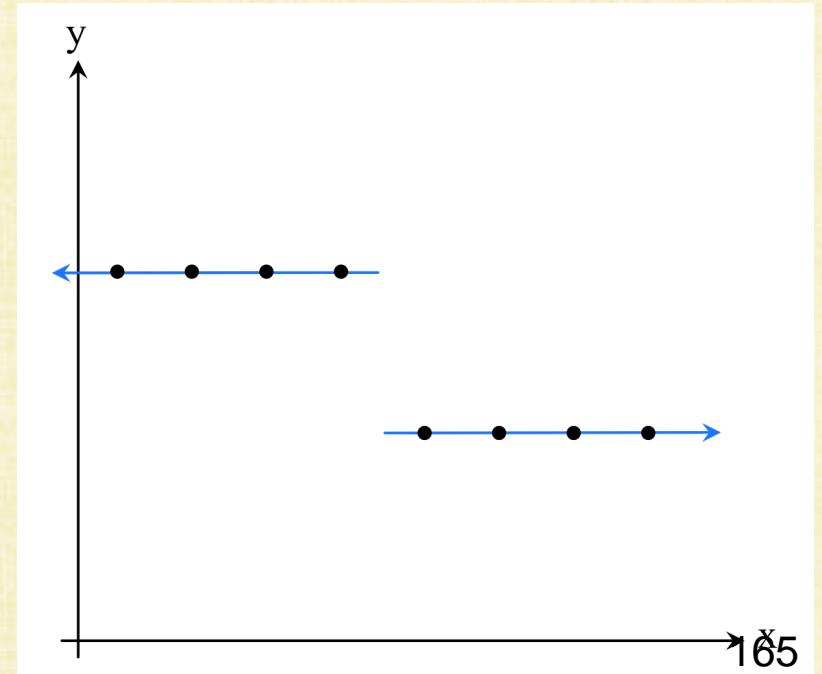
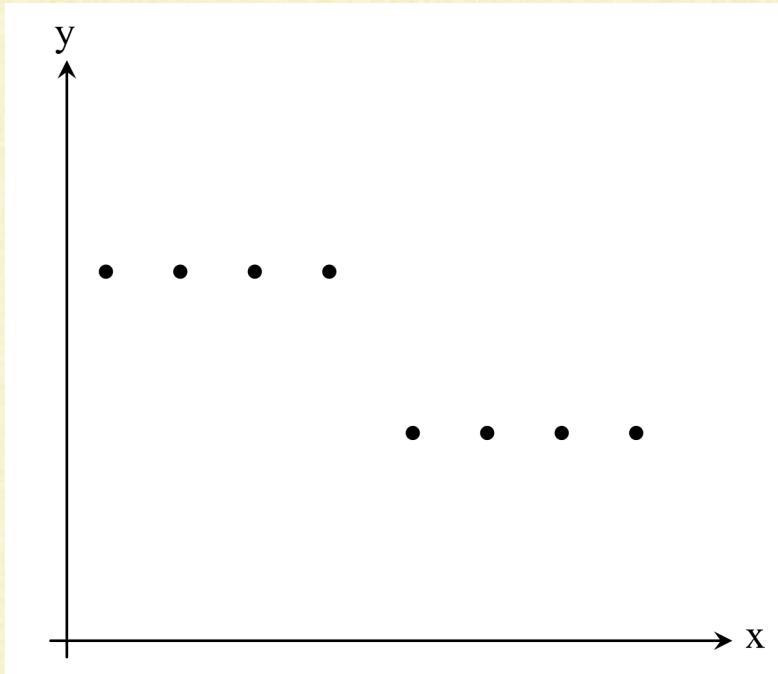
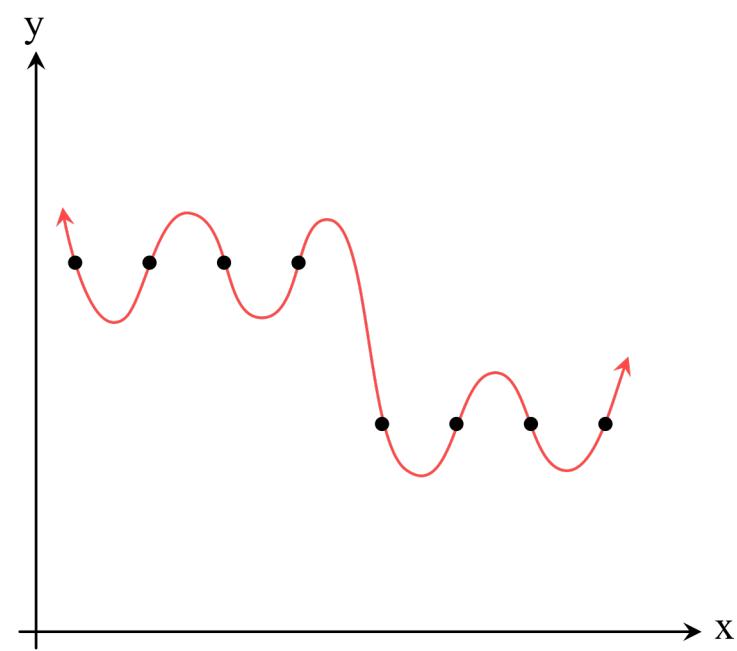
Well-Resolved Functions

- Regions of a function with less/more variation require lower/higher sampling rates



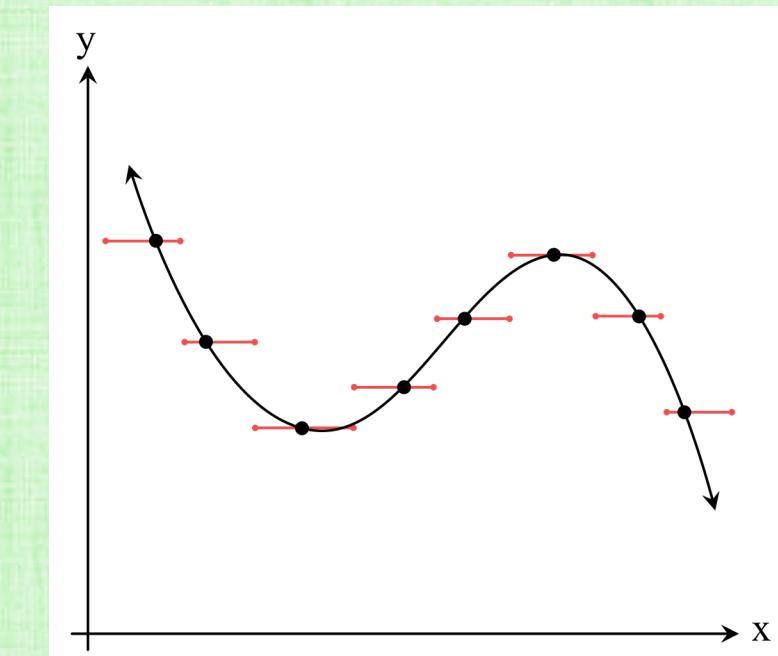
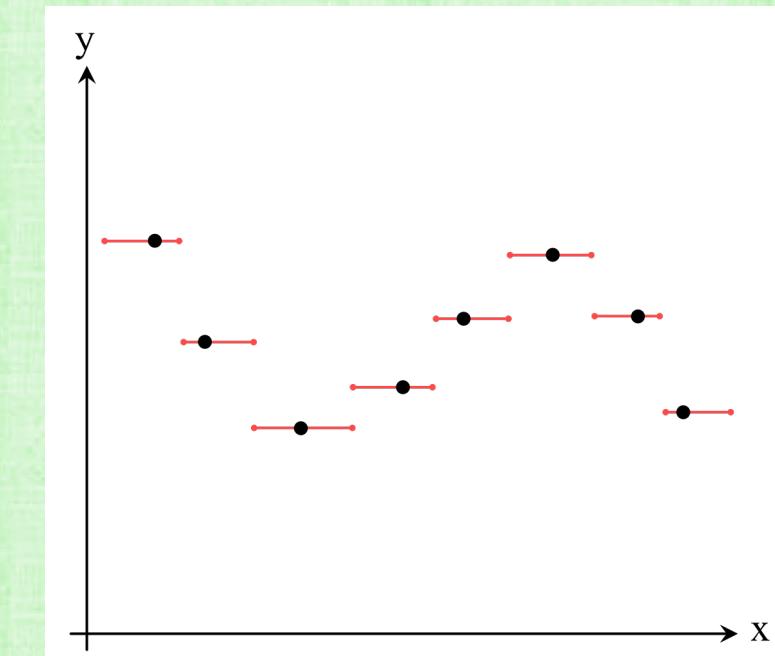
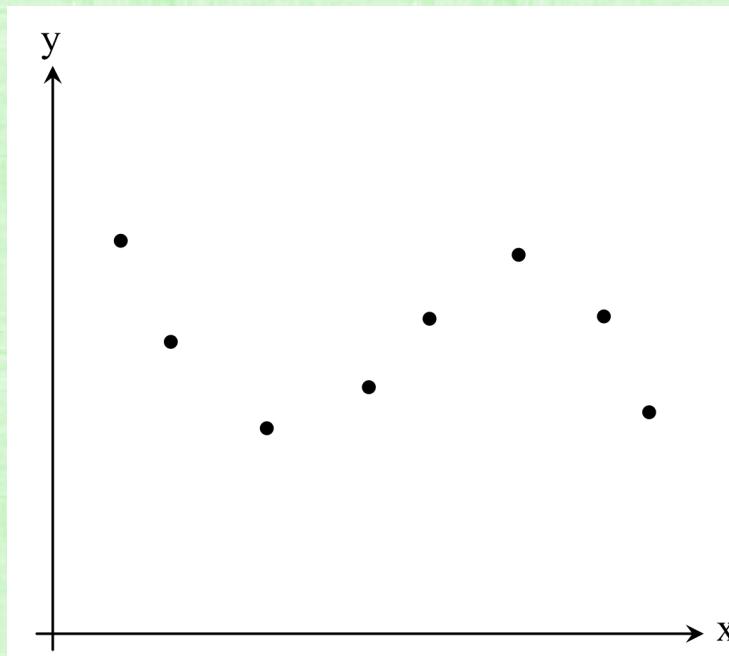
Piecewise Approximation

- Piecewise approximation enables the use of simpler models to approximate (potentially disjoint) subsets of data
 - ML/DL community: “sub-manifold” often means a coherent/smooth subset



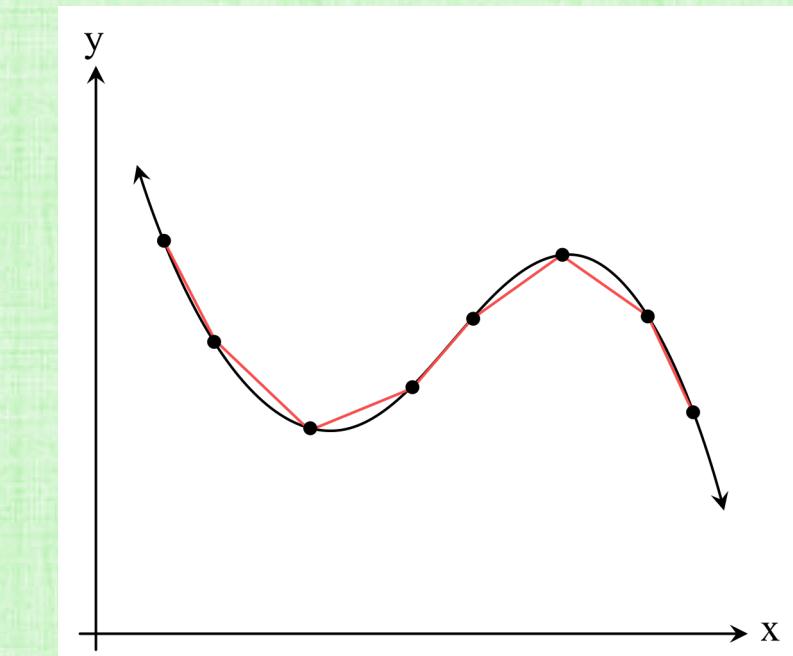
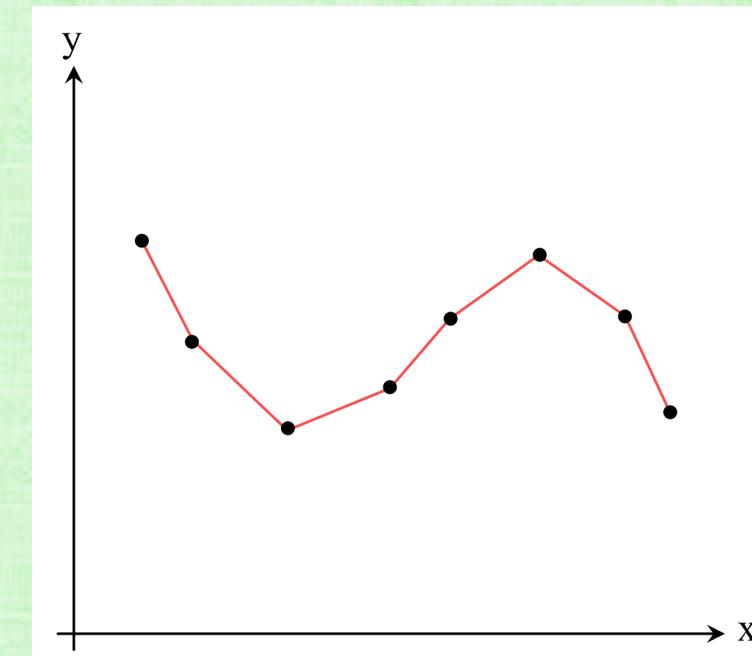
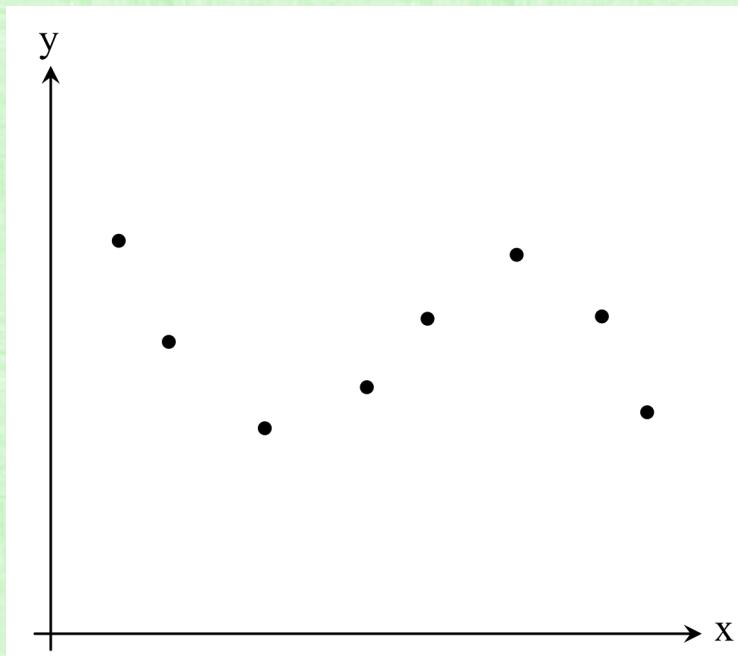
Piecewise Constant Interpolation

- Use the first term in the Taylor expansion (only): $f(x + h) \approx f(x)$
- Errors are $O(h)$, since $f(x + h) = f(x) + O(h)$
- Recall: nearest neighbor is also piecewise constant



Piecewise Linear Interpolation

- Use the first two terms in the Taylor expansion: $f(x + h) \approx f(x) + hf'(x)$
- Errors are $O(h^2)$, since $f(x + h) = f(x) + hf'(x) + O(h^2)$

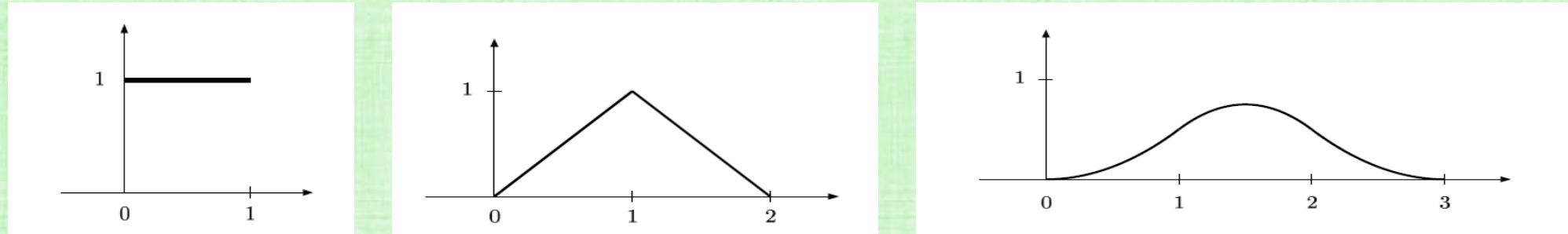


Higher Order Piecewise Interpolation

- Piecewise quadratic interpolation uses the first three terms in the Taylor expansion and has $O(h^3)$ errors
- Piecewise cubic interpolation uses the first four terms in the Taylor expansion and has $O(h^4)$ errors
- Recall: higher order interpolation becomes more oscillatory (i.e. overfitting)
 - These oscillations are also sometimes referred to as Gibbs phenomena

Cubic Splines

- Piecewise cubic splines are quite popular because of their ability to match derivatives across approximation boundaries
- B-splines – hierarchical family: ϕ_i^p is a piecewise polynomial of degree p
 - Piecewise constant: $\phi_i^0(x) = 1$ for $x \in [x_i, x_{i+1}]$ and 0 otherwise
 - A linear $w_i^p(x) = \frac{x-x_i}{x_{i+p}-x_i}$ increases the polynomial degree of ϕ^{p-1} to ϕ^p
 - Recursively: $\phi_i^{p+1}(x) = w_i^p(x)\phi_i^p(x) + (1 - w_{i+1}^p(x))\phi_{i+1}^p(x)$
 - Piecewise linear ϕ_i^1 , piecewise quadratic ϕ_i^2 , piecewise cubic ϕ_i^3 , etc.



2D Image Segmentation

- Divide an image's pixels into separate regions representing objects or groups of objects
- Traditional methods relied on clustering in color/space, graph-cut, edge detection, etc.
- More recently: train/use neural networks that can (hopefully) be driven more so by human perception/semantics
- Training examples:
 - Input: an image (all the pixel RGB values)
 - Output: labeling of all the pixels as to what group each corresponds to

(Example) Binary Output Labels

- Binary segmentation of an image using binary values
- E.g. 1 = dog, 0 = not dog



Input



Output

(Example) Integer Output Labels

- Multi-object segmentation with an integer for each object
- E.g. 1=cat, 2=dog, 3=human, 4=mug, 5=couch, 6=everything else



Input



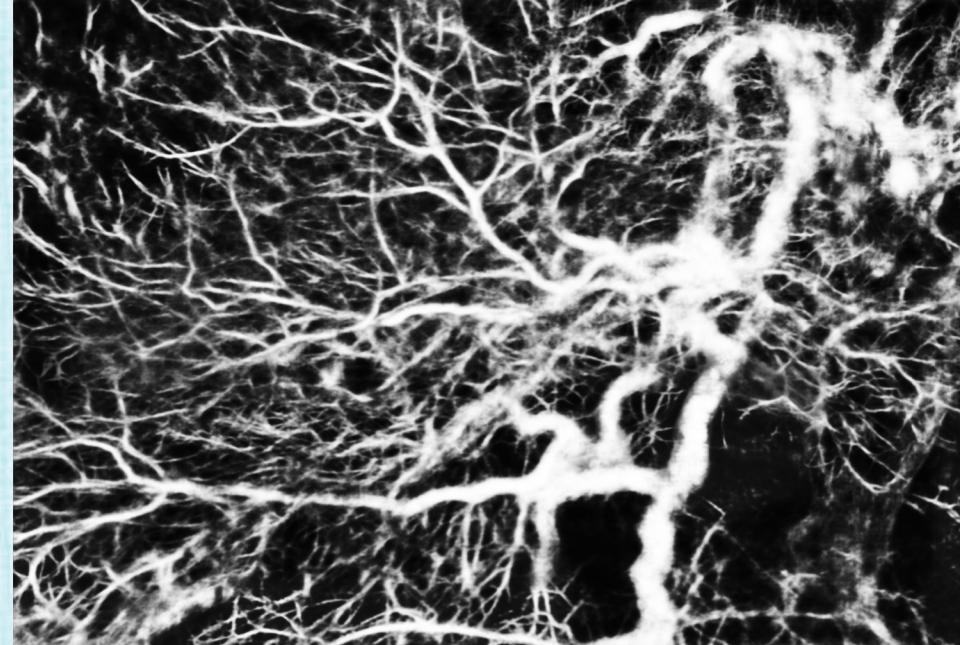
Output

(Example) Real Number Output Labels

- Probabilistic segmentation with real number values in [0,1]
- E.g. 1=tree branch, .8=probably a branch, .2=probably not a branch, etc.



Input



Output

Segmenting Botanical Trees (with a Network)

- Trees are large scale geometrically complex structures
- Branches severely occlude each other
- The images have limited pixel resolution of individual branches
- Even humans have a hard time ascertaining the correct topological structure from a single image/view
- Train a neural network to help

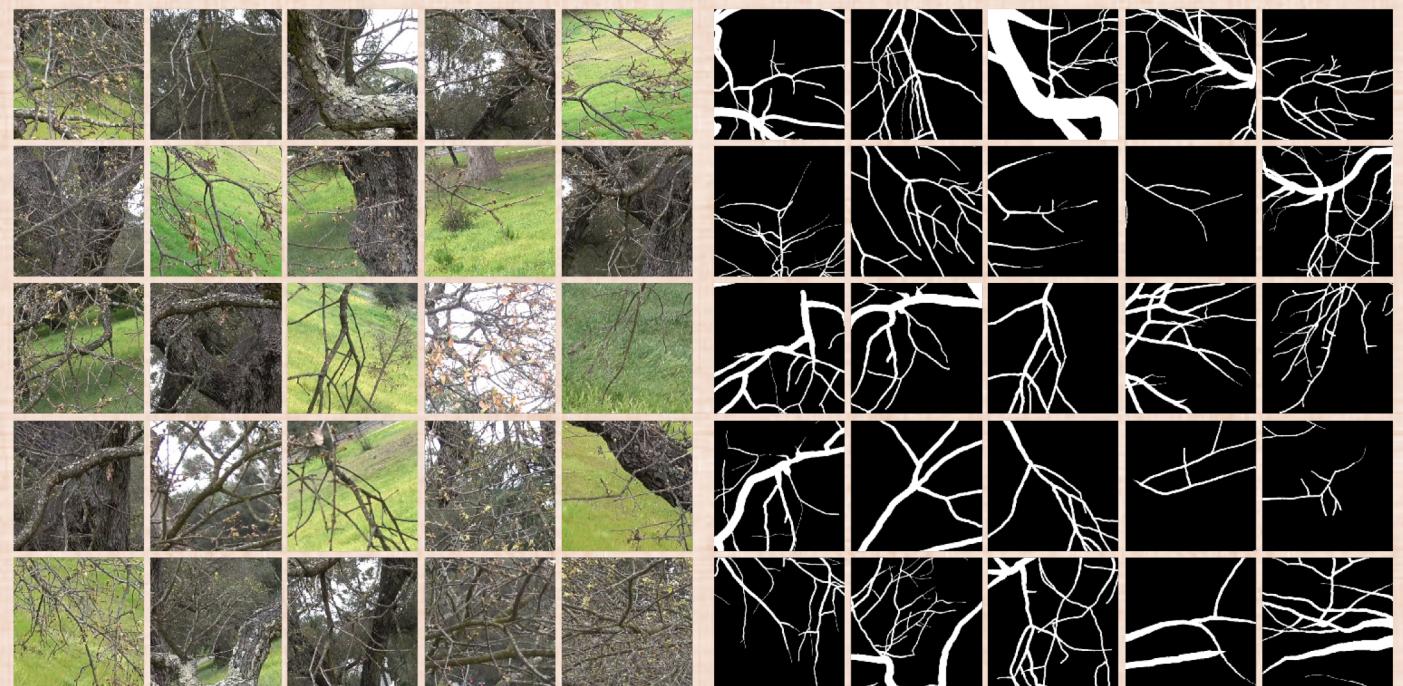
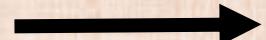
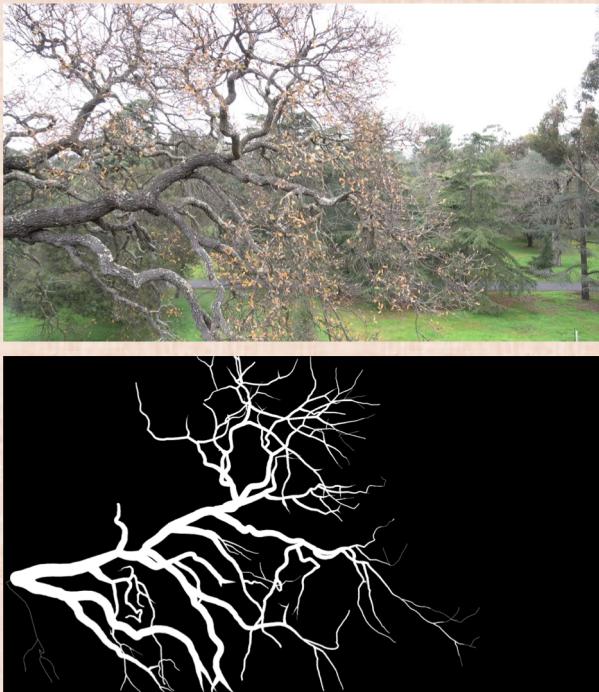
Constructing Training Data (By Hand)

- Begin with a dataset of labels created by people
- Draw lines and thicknesses on top of branches, and then flatten this information into a binary mask of the image



Constructing (More) Training Data

- Artificially increase the amount of training data by taking various image crops
- Also need to down-sample the resolution a bit (for the network)



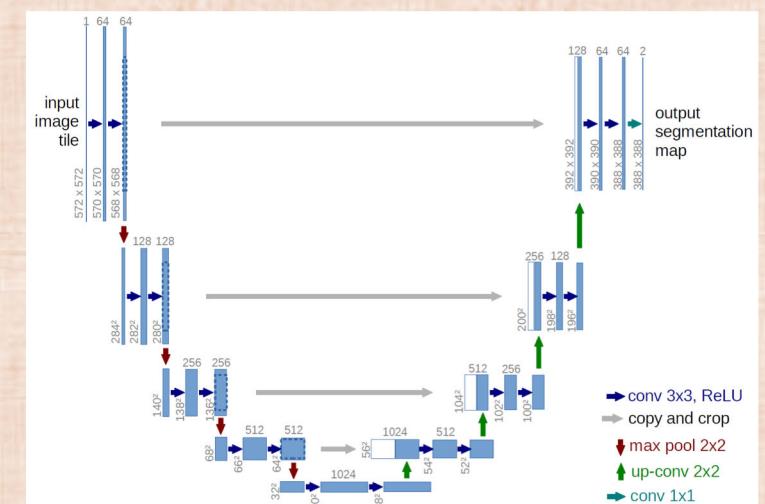
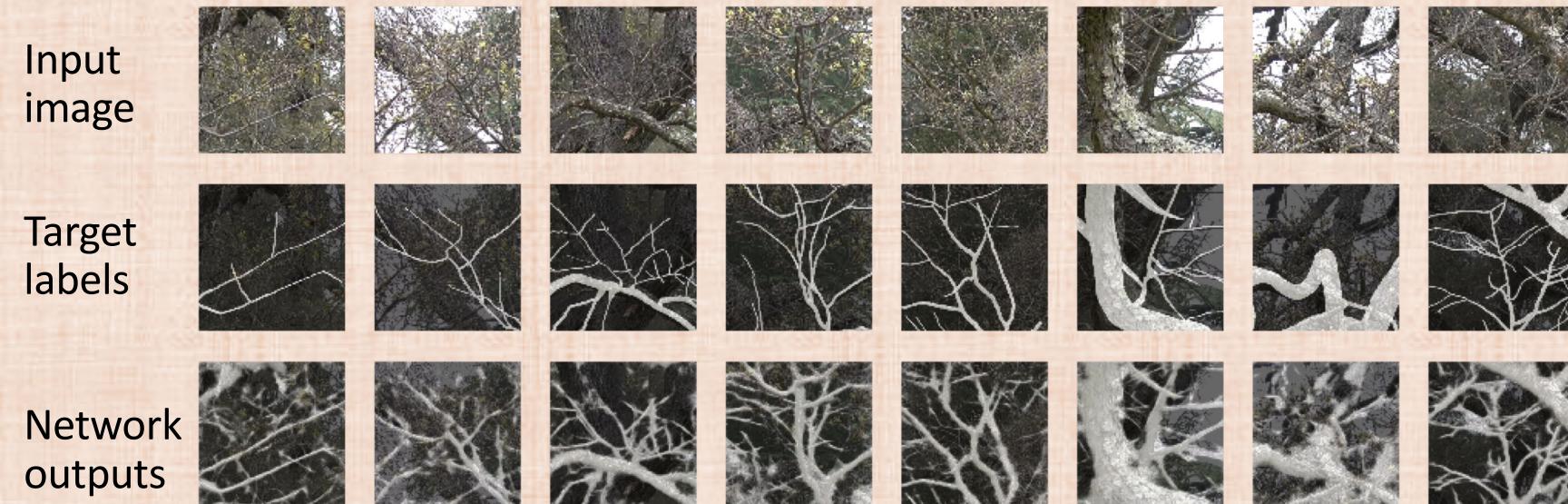
3840 pixels wide, 2160 pixels tall

512 pixels wide, 512 pixels tall

176

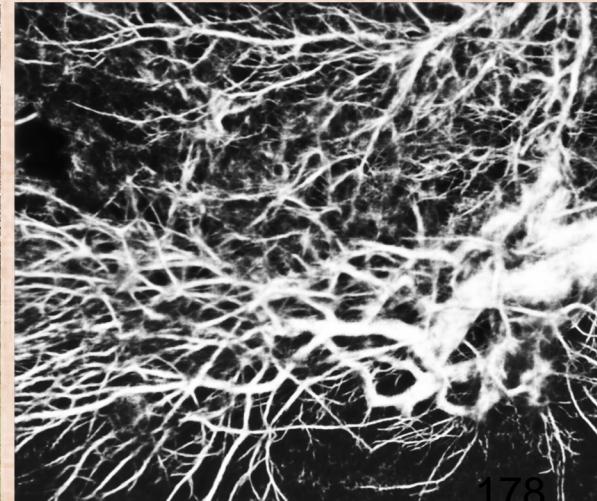
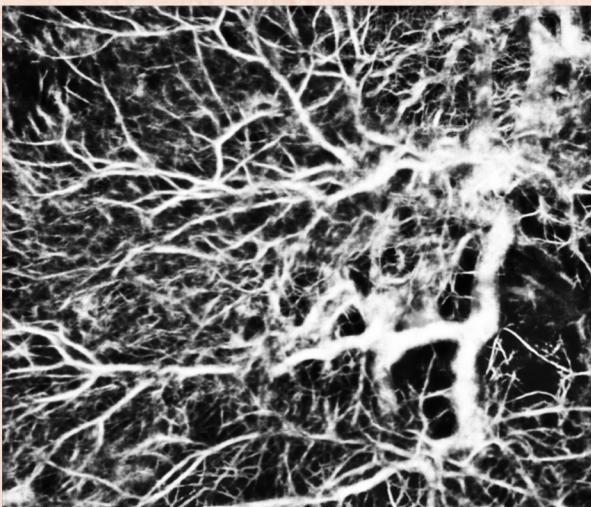
Training the Neural Network

- Find function parameters c such that the network function $f_c(x)$ gives minimal error on the training data (i.e. minimize network loss)
- Network should predict the known target labels (or close to it) from the input image



Network Inference/Prediction

- After training, use the resulting network function $f_{c_{trained}}(x)$ to infer/predict labels for new images that were not previously hand-labeled



Local Approximations

- Input images seem to be of two different types: either (1) branches over grass or (2) clusters of branches



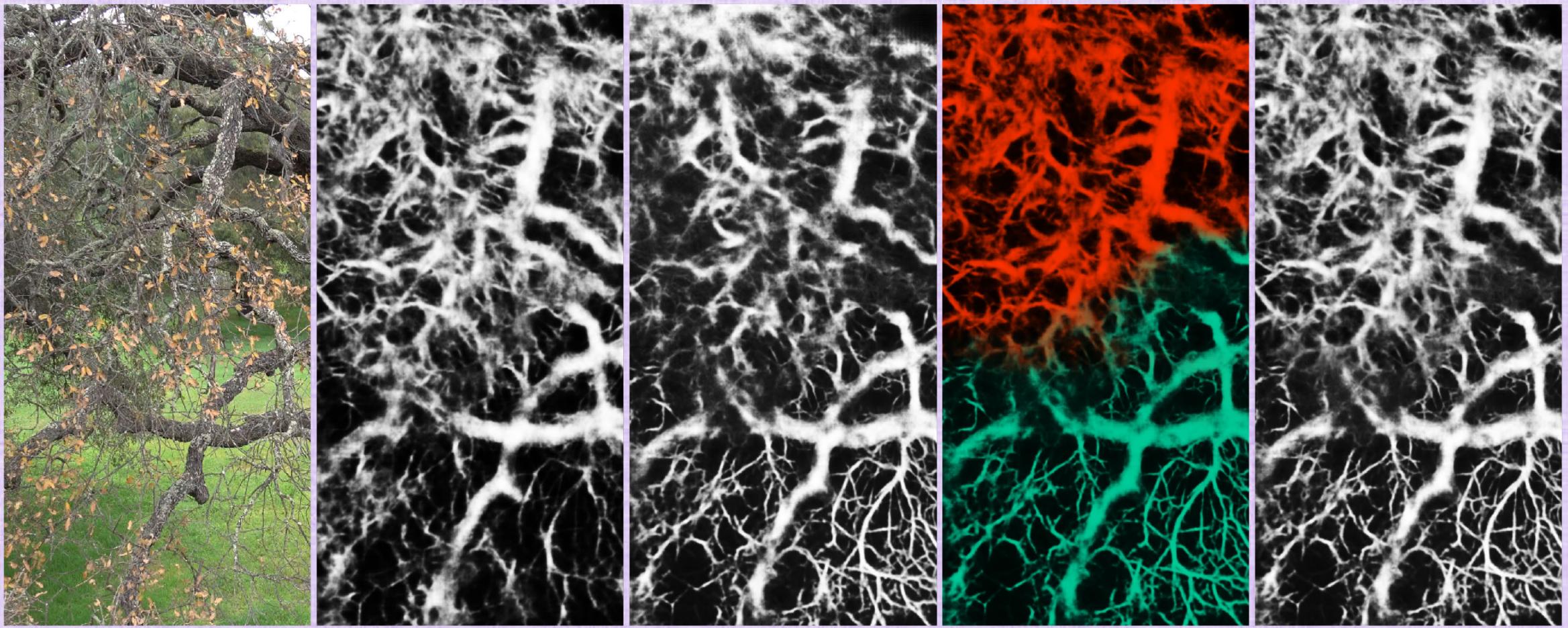
Local Approximation

- Network Construction:
 - Divide the training data into these two disparate groups
 - Train a separate network on each group of data: separate network architecture, separate trainable parameters, etc.
- Network Inference:
 - Given an input image, evaluate it separately on each network
 - Then combine the two results using the results that make the most sense locally in the image

Combining Inference Outputs

- To divide the training images into clusters, k-means clustering was done on hue and saturation
- To inference each pixel, take a small sample crop of the image around the pixel
- Compute hue/saturation values on the crop
- Find the distances from those hue/saturation values to the 2 cluster centers
- Interpolate the outputs from the 2 networks using those distances
- The closer a pixel is to a k-means cluster, the more weight that cluster's network inference/prediction is given

Example



Input

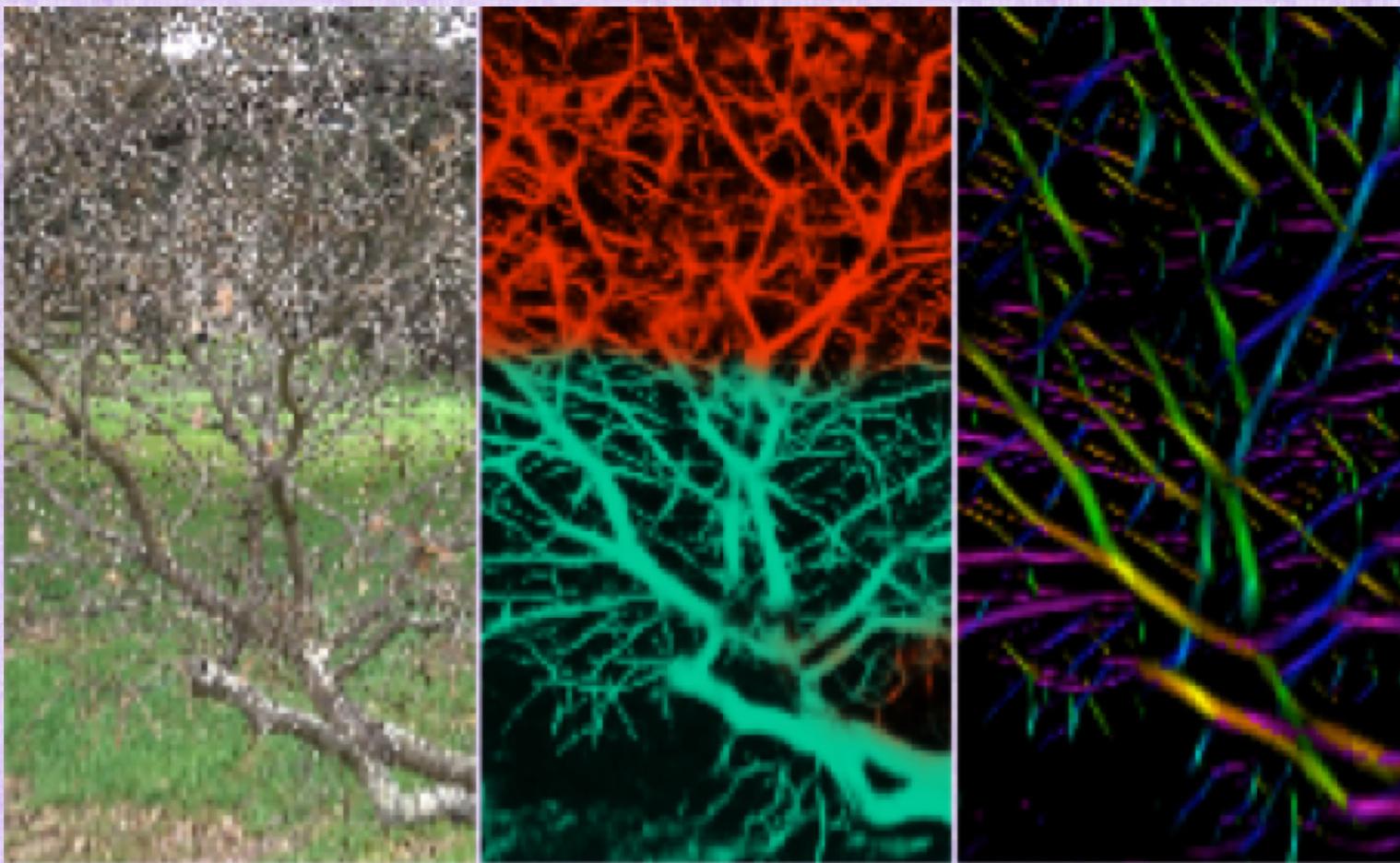
Network 1

Network 2

Combine

Final Result

Aside: Branch Estimation



Unit 7

Curse of Dimensionality

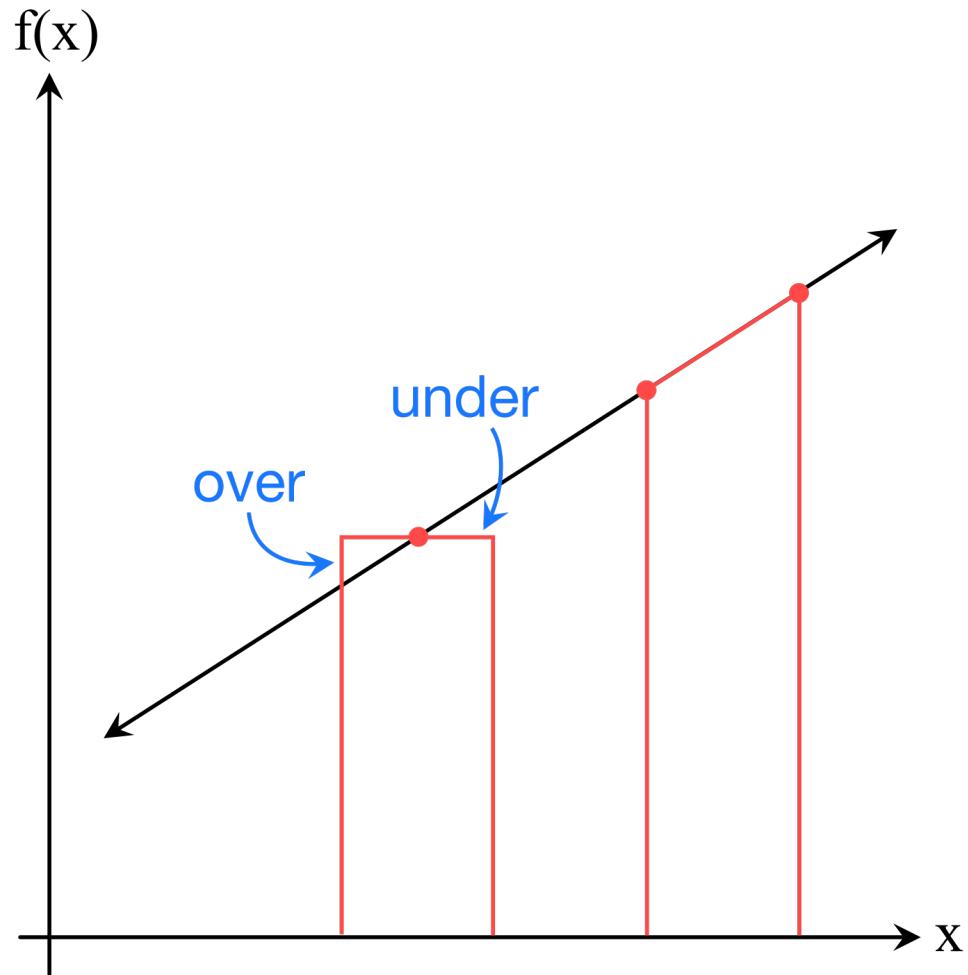
Numerical Integration (Quadrature)

- Approximate $\int_{x_L}^{x_R} f(x)dx$ numerically
- Break up $[x_L, x_R]$ into subintervals, and consider each subinterval separately
- On each subinterval:
 - Reconstruct the function
 - Analytically find the area under the reconstructed curve
- These two steps can be combined in various ways (for efficiency)
- f is often not explicitly known
- I.e., often only have access to output values $f(x_i)$ given input values x_i
- In addition, could be expensive to evaluate $f(x_i)$, especially when it requires running code (e.g. neural network inference)

Newton-Cotes Quadrature

- On each subinterval, choose p equally spaced points and use $p - 1$ degree polynomial interpolation to reconstruct the function and approximate the integral
- Obtains the exact solution when f is a degree $p - 1$ polynomial (as expected)
- When the number of points p is odd, symmetric cancellation gives the exact solution on a degree p polynomial (1 degree higher than expected)

Symmetric Cancellation



- When $p = 2$ points, the 1st degree piecewise linear approximation integrates piecewise linear functions exactly
- When $p = 1$ point, the 0th degree piecewise constant approximation (also) integrates piecewise linear function exactly
 - Note cancellation of under/over approximations in the figure

Newton-Cotes Quadrature (Examples)

- Consider a total of m intervals
- Piecewise constant approximation ($p = 1$ point) uses a total of m points to integrate piecewise linear functions exactly
- Piecewise linear approximation ($p = 2$ points) uses a total of $m + 1$ points to integrate piecewise linear functions exactly
 - points on the boundary between intervals are used for both intervals
- Piecewise quadratic approximation ($p = 3$ points) uses a total of $2m + 1$ points to integrate piecewise cubic functions exactly
- Piecewise cubic approximation ($p = 4$ points) uses a total of $3m + 1$ points to integrate piecewise cubic functions exactly

Local and Global Error

- Degree p polynomial reconstruction captures the Taylor expansion terms up to and including $\frac{h^p}{p!} f^{(p)}(x)$, with $O(h^{p+1})$ errors
- This $O(h^{p+1})$ error in the height of the function multiplied times the $O(h)$ width of the interval gives per interval local area error of $O(h^{p+2})$
- The total number of intervals is $\frac{x_R - x_L}{O(h)} = O\left(\frac{1}{h}\right)$, so the total global error is $O\left(\frac{1}{h}\right) O(h^{p+2}) = O(h^{p+1})$
- Doubling the number of intervals halves their size leading to $\left(\frac{1}{2}\right)^{p+1}$ as much error, which is denoted an order of accuracy of $p + 1$

Newton-Cotes Quadrature (Examples)

- Midpoint Rule: $\sum_i h_i f(x_i^{mid})$
 - 1 point, piecewise constant, exact for piecewise linear, 2nd order accurate
- Trapezoidal Rule: $\sum_i h_i \frac{f(x_i^{left}) + f(x_i^{right})}{2}$
 - 2 points, piecewise linear, exact for piecewise linear, 2nd order accurate
- Simpson's Rule: $\sum_i h_i \frac{f(x_i^{left}) + 4f(x_i^{mid}) + f(x_i^{right})}{6}$
 - 3 points, piecewise quadratic, exact for piecewise cubic, 4th order accurate

Gaussian Quadrature

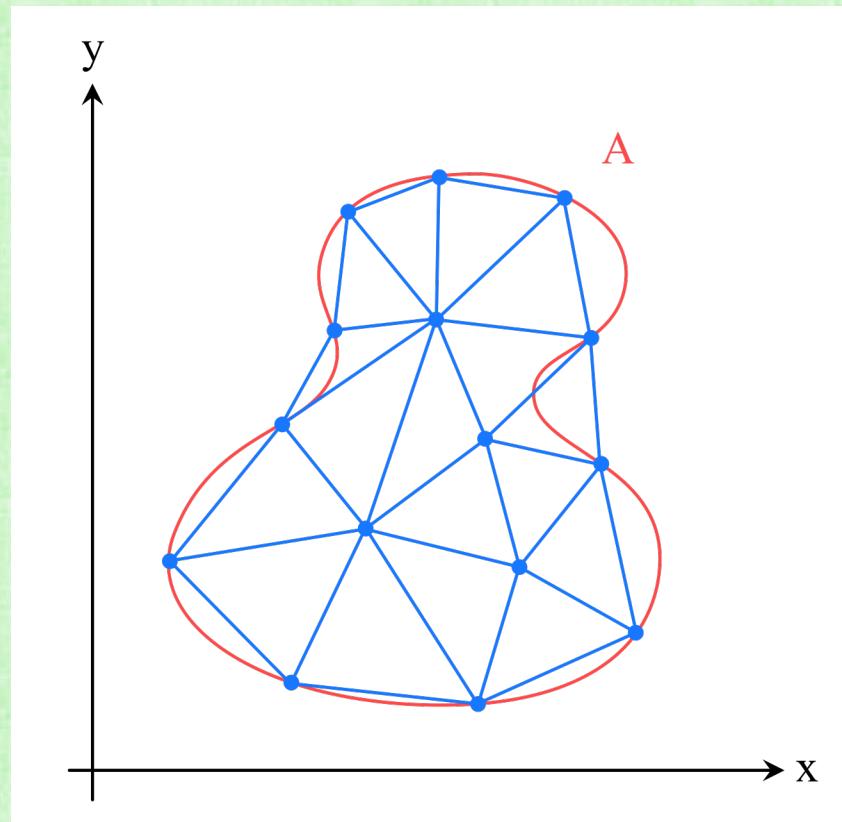
- Use p optimally chosen points to obtain a method that is exact on degree $2p - 1$ polynomials, and thus has an order of accuracy of $2p$
- For example: $\sum_i h_i \frac{f\left(x_i^{mid} - \frac{h_i}{2\sqrt{3}}\right) + f\left(x_i^{mid} + \frac{h_i}{2\sqrt{3}}\right)}{2}$
- 2 points, piecewise cubic, exact for piecewise cubic, 4th order accurate
- Same accuracy as Simpson's 3 point rule
 - Simpson has 1 point on the boundary, so only $2m + 1$ total points are required
 - That is, Gaussian quadrature only saves 1 point in total ($2m$ total points)

Two Dimensions

- $\iint_A f(x, y)dA$ where sub-regions dA of area A are considered separately
- When A is rectangular, it can be broken into sub-rectangles and addressed dimension-by-dimension using 1D techniques
- When A is more interesting, triangle sub-regions can be used to approximate it
- The difference between A and its approximation leads to a new source of error not seen in 1D (where interval boundaries are simply just 2 endpoints)

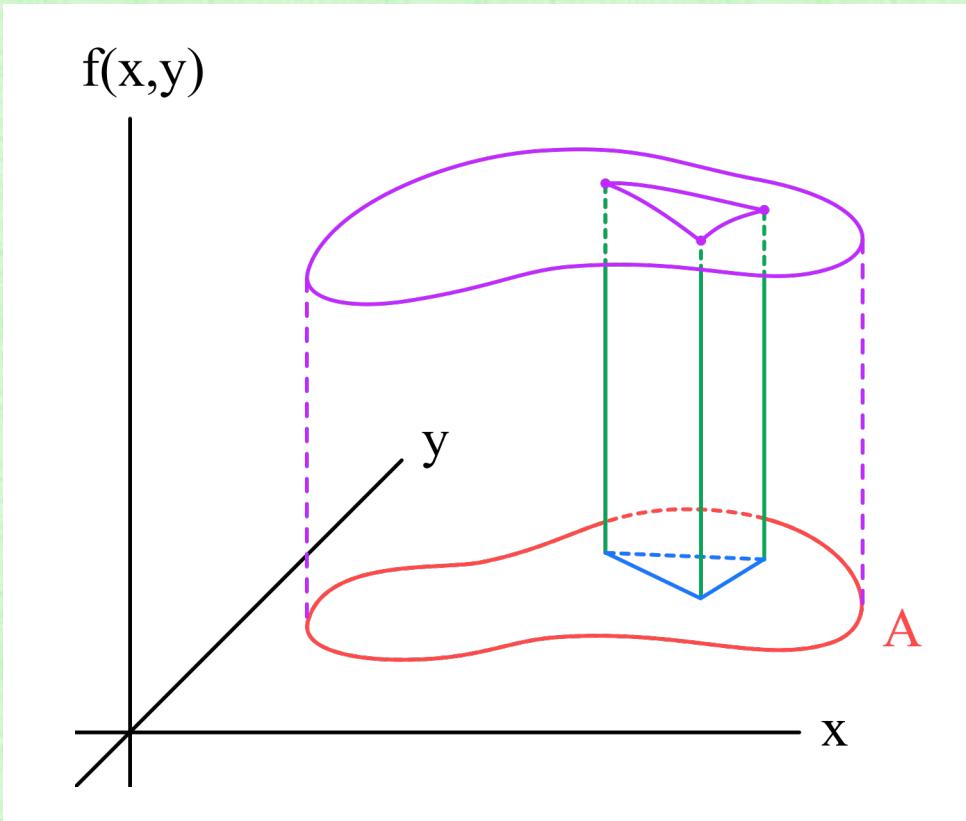
Domain Approximation Errors

- The difference between A and its approximation via triangles leads to a new source of error in the integral (there is missing/extra area)



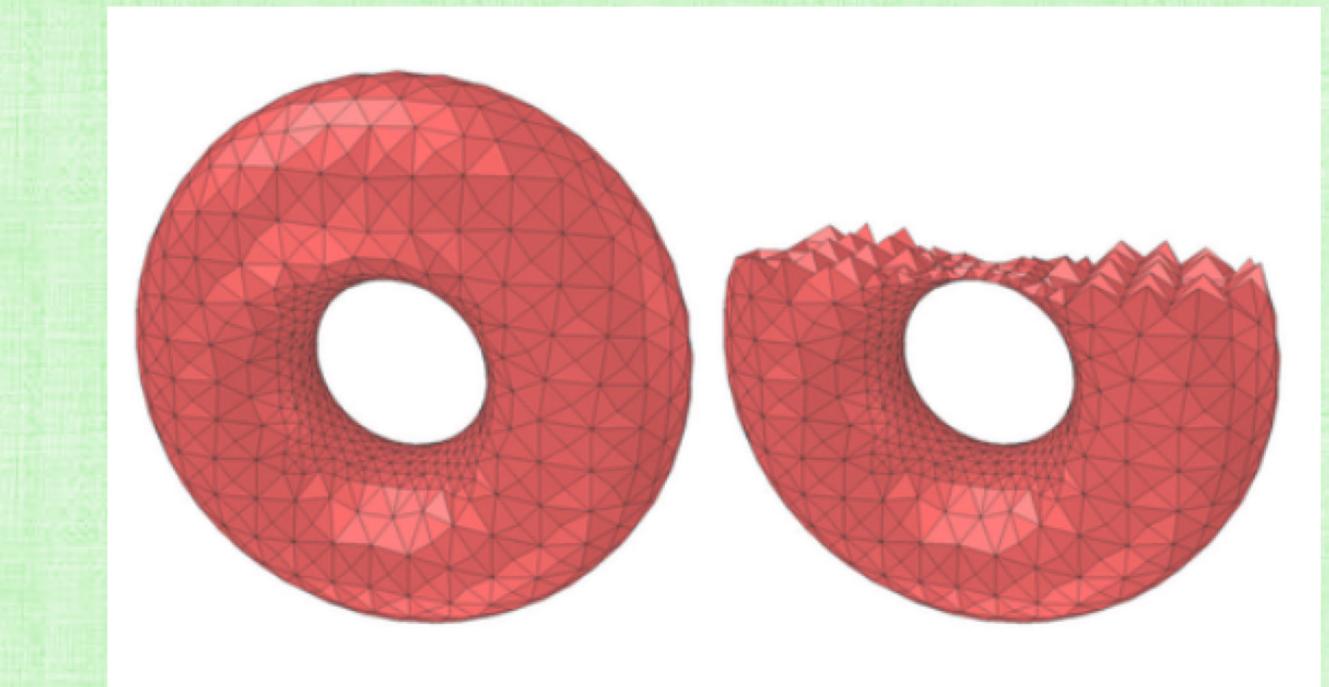
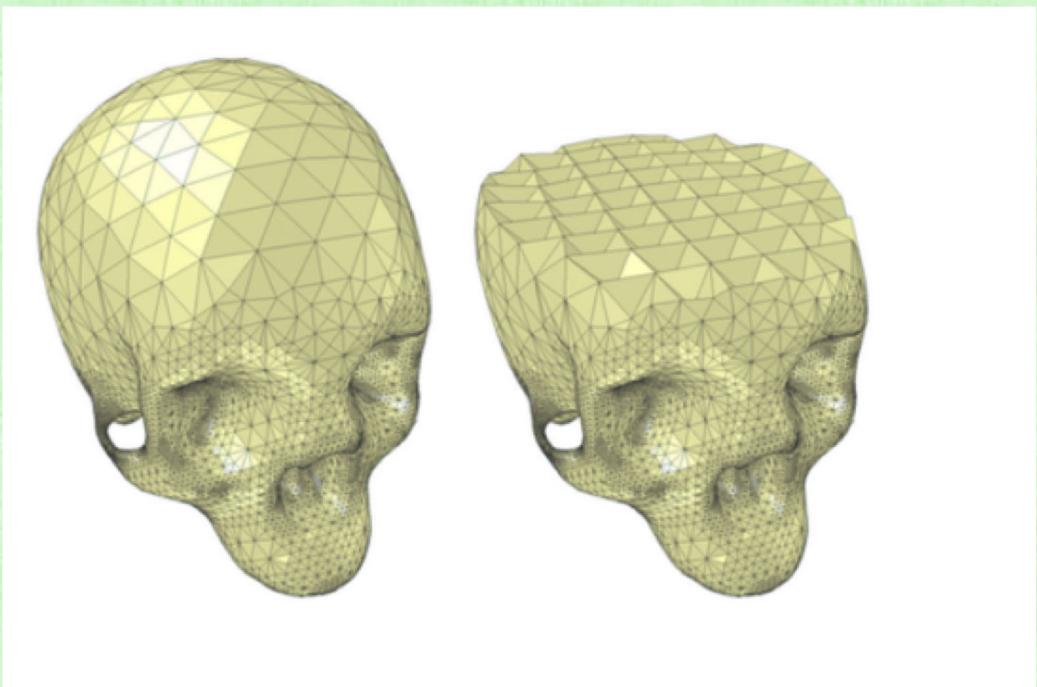
Integrating Sub-regions

- Each triangle sub-region utilizes optimally chosen Gaussian quadrature points to compute sub-volumes



Three Dimensions

- $\iint_V f(x, y, z) dV$ where tetrahedral sub-regions dV of volume V are considered separately (with Gaussian quadrature points)



Curse of Dimensionality

- Consider a 1st order accurate method
- 1D: doubling the number of intervals cuts the error in half ($2x$ work = $\frac{1}{2}$ error)
- 2D: halving interval size requires 4 times the rectangles/triangles ($4x$ work = $\frac{1}{2}$ error)
- 3D: halving interval size requires 8 times the cubes/boxes/tets ($8x$ work = $\frac{1}{2}$ error)
- 4D: $16x$ work = $\frac{1}{2}$ error, 5D: $32x$ work = $\frac{1}{2}$ error, etc.
- Cutting error by a factor of 4 in 5D takes $32^2=1024x$ work
- Cutting error by a factor of 8 in 5D takes $32^3=32,768x$ work
- If the original code took 1 sec to run in 5D, cutting error by a factor of 8 takes 9 hours
- And cutting error by a factor of 16 takes 12 days
- And cutting the error by a factor of 32 takes over a year....

Yep, you're cursed

Curse of Dimensionality

- Consider a 2nd order accurate method
- In 1D/2D/3D/4D/5D/etc. halving the intervals gives 4 times less error
- Cutting error by a factor of 4 in 5D (still) takes 32x work
- If the original code (still) took 1 sec to run in 5D, cutting error by a factor of 16 takes only 17 min (instead of 12 days)
- But cutting error by a factor of 1024 (3 decimal places more accuracy) takes over a year...
- In 10D, cutting error by a factor of 4 takes 1024x work
- Second order is better than first, but still intractable in higher dimensions
- Moreover, it's difficult/impossible to construct higher order methods in higher dimensions (and overfitting is a concern too)

Conclusion

- Newton-Cotes style approaches are only practical for 1D/2D/3D
 - or 1D/2D/3D + time

A (fun) Example

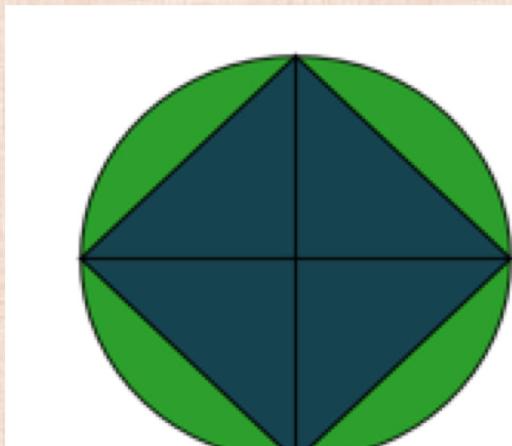
- Consider approximating $\pi = 3.1415926535 \dots$
- Use a compass to construct a circle with radius = 1
- Since $A = \pi r^2$, the area of the circle is π
- Setting $f(x, y) = 1$ gives $\iint_A f(x, y)dA = \pi$



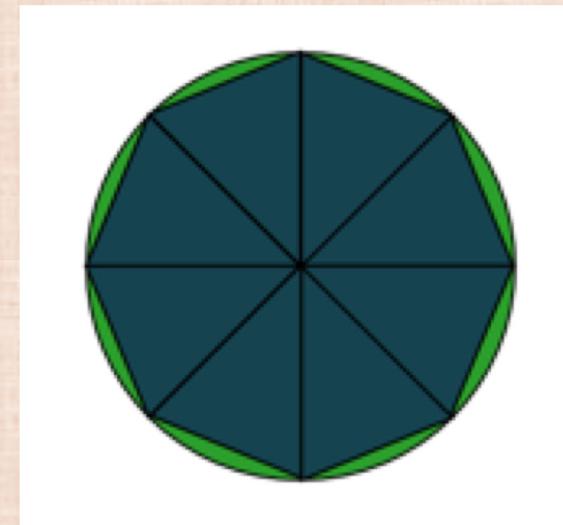
$$A = \pi$$

Newton-Cotes Approach

- Inscribe triangles inside the circle
- The function $f(x, y) = 1$ translates into computing the area of each triangle
- The difference between A and its approximation with triangles leads to errors



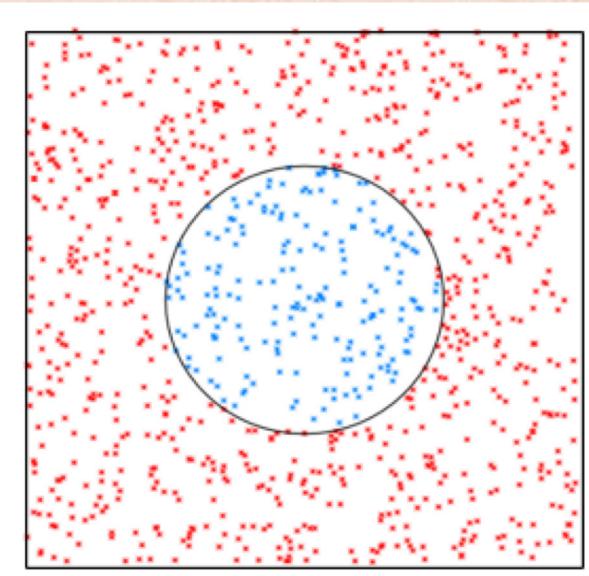
$$\pi \approx 2$$



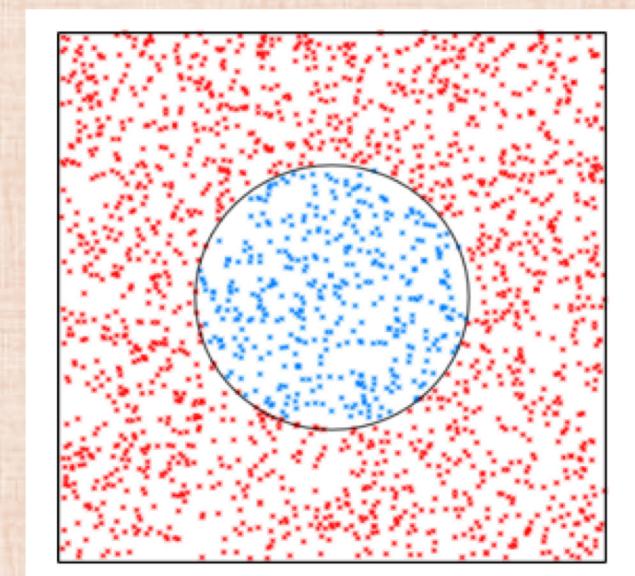
$$\pi \approx 2.8284$$

Monte Carlo Approach

- Construct a square with side length 4 containing the circle
- Randomly generate N points in the square, and color points inside the circle blue
- Since $\frac{A_{circle}}{A_{box}} = \frac{\pi}{16}$, one can approximate $\pi \approx 16 \left(\frac{N_{blue}}{N_{total}} \right)$



$$\pi \approx 3.136$$



$$\pi \approx 3.1440$$

Monte Carlo Methods

- Typically used in higher dimensions (5D or more)
- Random (pseudo-random) numbers generate sample “points” that are multiplied by element “size” (e.g. length, area, volume)
- Error decreases like $\frac{1}{\sqrt{N}}$ where N is the number of samples (1/2 order accurate)
 - E.g. 100 times more points are needed to gain one more digit of accuracy
- Very slow convergence, but independent of the number of dimensions!
- Not competitive for lower dimensional problems, but the only tractable alternative for higher dimensional problems

Machine Learning Implications

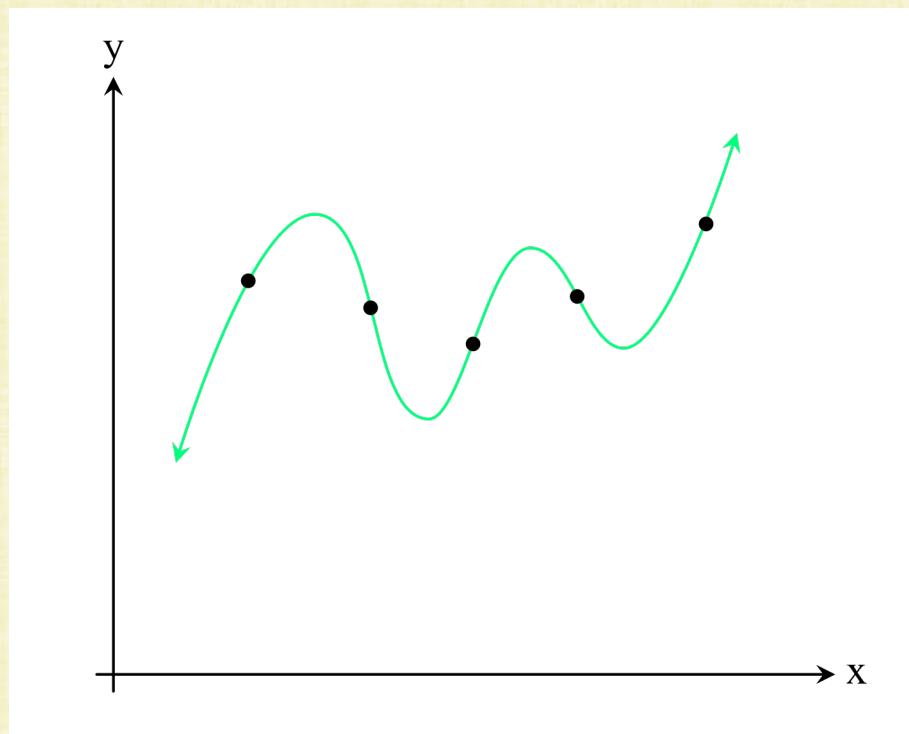
- Consider $y = f(x)$ where $x \in R^n$ with large n
- Newton-Cotes style approaches first do polynomial interpolation, and then analytically integrate the result
- The curse of dimensionality occurs because of the sheer number of points required to construct polynomial functions in higher dimensions
- The same sentiments hold true when constructing $y = f(x)$ for interpolation/inference, i.e. a higher dimensional x is intractable
- Thus, Monte Carlo approaches are far more efficient!
- This is a major reason for close collaborations between ML/DL and Statistics departments, as compared to classical engineering (which 3D models the physical world, and as such has closer ties to Applied Math)

Unit 8

Least Squares

Recall: Polynomial Interpolation

- Given m data points, can draw a unique $m - 1$ degree polynomial through them
 - As long as they're not degenerate, like 3 points on a line, which we'll discuss later

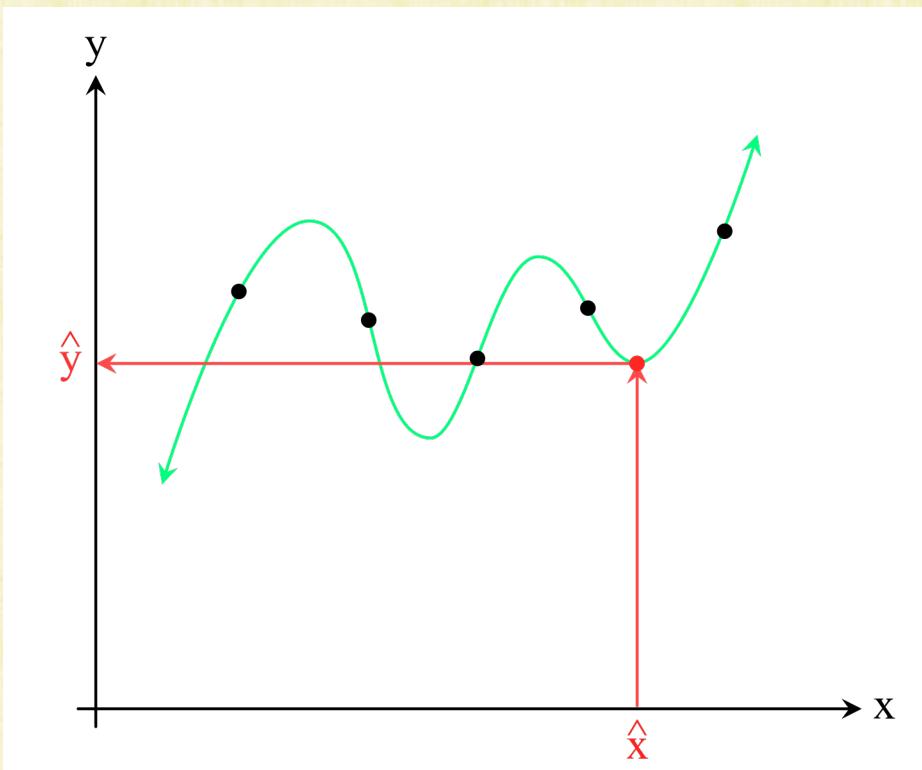


Recall: Basis Functions

- Given basis functions ϕ and unknowns c : $y = c_1\phi_1 + c_2\phi_2 + \cdots + c_n\phi_n$
- Monomial basis: $\phi_k(x) = x^{k-1}$
- Lagrange basis: $\phi_k(x) = \frac{\prod_{i \neq k} x - x_i}{\prod_{i \neq k} x_k - x_i}$
- Newton basis: $\phi_k(x) = \prod_{i=1}^{k-1} x - x_i$
- Write a (linear) equation for each point, and put into matrix form: $Ac = y$
- Monomial/Lagrange/Newton basis all give the same polynomial

Recall: Overfitting

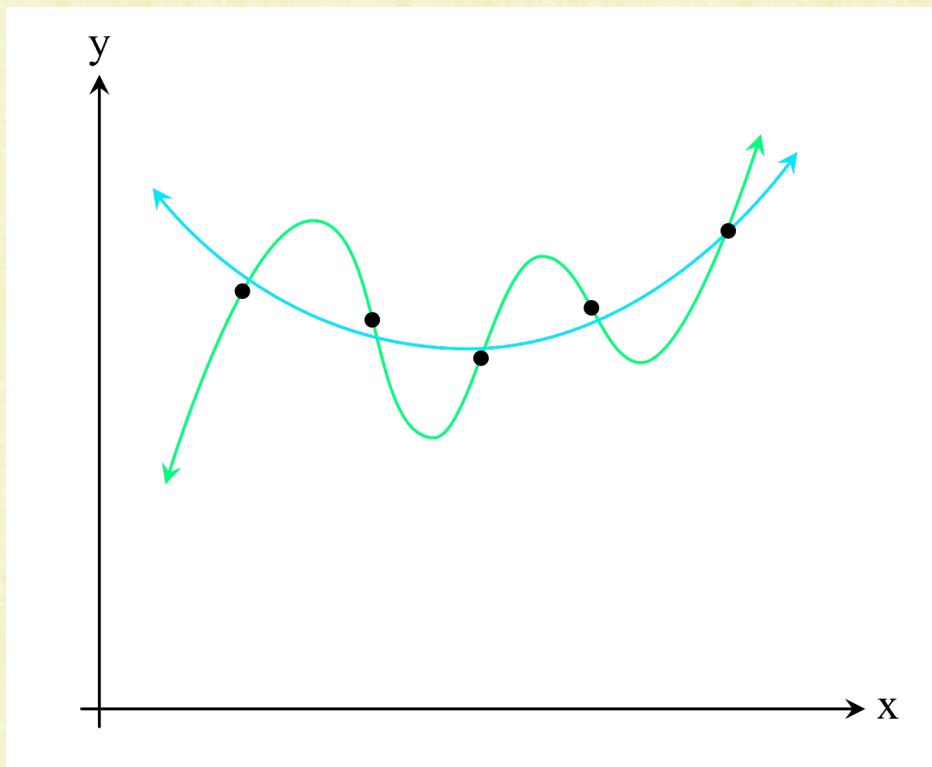
- Given new input \hat{x} , the interpolant is used to infer/predict \hat{y} (that may be far from what is expected)



- Interpolating polynomials are smooth (continuity of the function/derivatives)
- Thus, they overshoot in between data points in order to smoothly pass through all of them
- Forcing a polynomial to fit every data point is called overfitting (overly fitting to the data)
- It results in inference/predictions that vary too wildly from the training data

Recall: Regularization

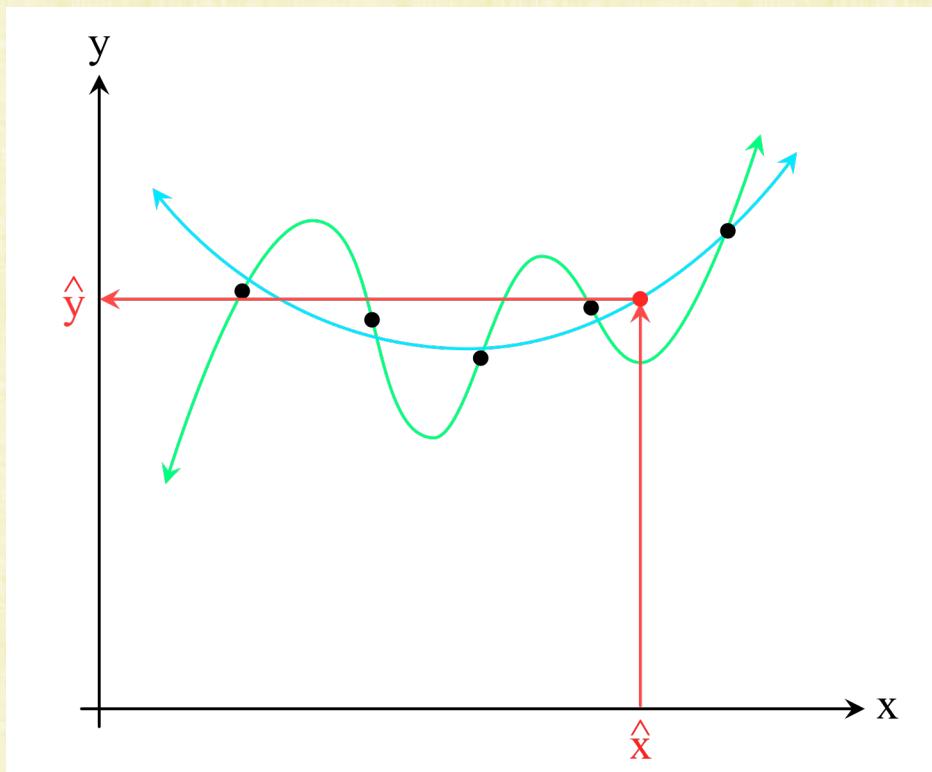
- Using a lower order polynomial that doesn't/can't fit the data points provides some regularization



- A regularized interpolant contains intentional errors, missing some/all the data points
- However, this (hopefully) makes the function more predictable/smooth between data points
- Moreover, the data points themselves may contain noise/error, so it's not clear they should be interpolated exactly

Recall: Regularization

- Given \hat{x} , the regularized interpolant infers/predicts a more reasonable \hat{y}



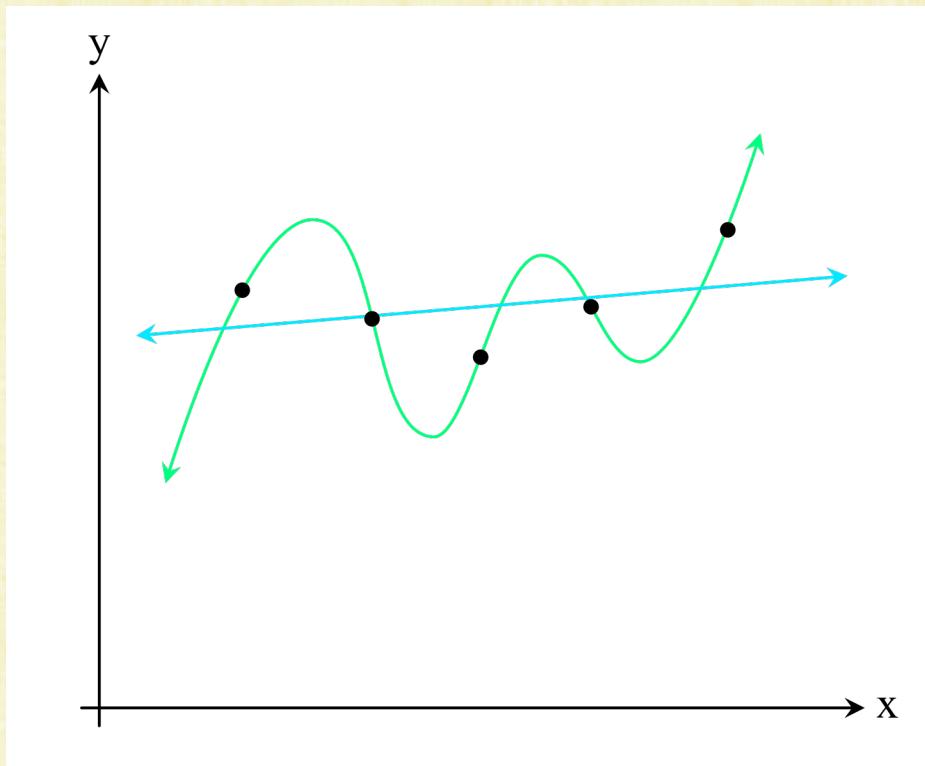
- Trade-off between sacrificing accuracy on interpolation of the input data, and obtaining better accuracy on inference/prediction for new inputs

Eliminating Basis Functions

- Consider $Ac = y$:
 - Each row of A evaluates all n basis functions ϕ_k on a single point x_i
 - Each column of A evaluates all m points x_i on a single basis function ϕ_k
- Regularize by reducing the number of basis functions (and thus the degree of the polynomial)
- Otherwise the process proceeds as usual
 - I.e., write a (linear) equation for each point, and put into matrix form: $Ac=y$
- When there are more points than basis functions, there are more rows than columns (and the matrix is tall/rectangular)
- The matrix still has full (column) rank, as long as the basis functions are linearly independent (and the data isn't degenerate)

Recall: Underfitting

- Using too low an order polynomial leads to missing the data by too much



- A linear function doesn't capture the essence of this data as well as a quadratic function does
- Choosing too simple of a model function or regularizing too much prevents proper interpolation of the data

Tall (Full Rank) Matrices

- Let A be a size $m \times n$ tall (i.e. $m > n$) matrix with full (column) rank (i.e. rank n)
- Since there are only n entries in each row, the rows span at most an n dimensional space (so at least $m - n$ rows are linear combinations of others)
- That is, A contains $m - n$ extra unnecessary equations (that are linear combinations of others)
- Thus, A could be reduced to n equations (and size $n \times n$) without losing any information (**always true, but doesn't consider the right hand side**)
- The SVD ($A = U\Sigma V^T$) illustrates this, as the last $m - n$ rows of Σ are identically zero
- The last $m - n$ columns in U are hit by these zeros, and thus not in the range of A

Recall: Example

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix} =$$

$$\begin{pmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{pmatrix} \begin{pmatrix} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{pmatrix}$$

- Singular values are 25.5 , 1.29 , and 0
- Singular value of 0 indicates that the matrix is rank deficient
- The rank of a matrix is equal to its number of nonzero singular values

Recall: Example

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix} =$$
$$\left(\begin{array}{cccc} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{array} \right) \left(\begin{array}{ccc} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \end{array} \right) \left(\begin{array}{ccc} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{array} \right)$$

- The 3D space of vector inputs can only span a 3D subspace of R^4
- The last (green) column of U represents the unreachable dimension, orthogonal to the range of A , and is always multiplied by 0
- One can delete this column and the associated portion of Σ and still obtain a valid factorization

Solving Linear Systems (the right hand side)

- $Ac = b$ becomes $U\Sigma V^T c = b$ or $\Sigma(V^T c) = (U^T b)$ or $\Sigma \hat{c} = \hat{b}$
- Solve $\Sigma \hat{c} = \hat{b}$ by dividing the entries of \hat{b} by the singular values σ_k , then $c = V\hat{c}$
- The last $m - n$ equations are identically zero on the left, and should be identically zero on the right as well
 - E.g. $\begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{c}_1 \\ \hat{c}_2 \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{pmatrix}$ requires $\hat{b}_3 = 0$ in order to have a solution
 - The last $m - n$ columns in U are not in the range of A , so b must be in the span of the first n columns of U in order for the problem to have a solution

False Statements

- Reasoning with a false statement leads to infinitely more false statements:

$$\begin{aligned} a &= b \\ a^2 &= ab \\ a^2 - b^2 &= ab - b^2 \\ (a + b)(a - b) &= b(a - b) \\ a + b &= b \\ b + b &= b \\ b(1 + 1) &= b(1) \\ 2 &= 1 \end{aligned}$$

- Don't make false statements!

False Statements

- Reasoning with a false statement leads to infinitely more false statements:

$$\begin{aligned} Ac &= b \\ A^T Ac &= A^T b \\ c &= (A^T A)^{-1}(A^T b) \end{aligned}$$

Is it? Is it really?

- Don't make false statements!
- A mix of false/true statements makes it difficult to keep track of what is and what is not true

False Statements

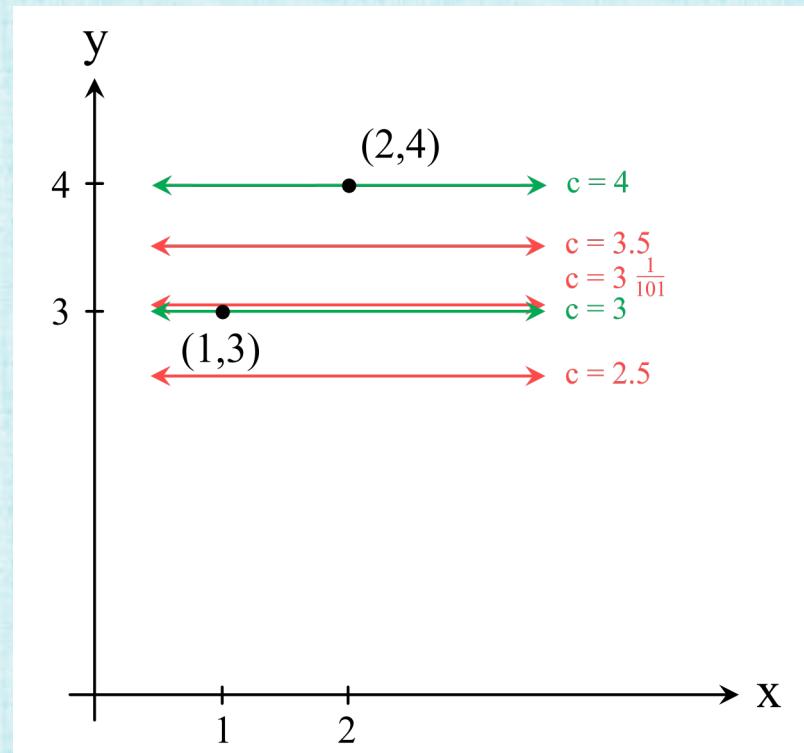
- Consider a very simple $Ac = b$ given by: $\begin{pmatrix} 1 \\ 1 \end{pmatrix}(c) = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$
- This contains the equations $c = 3$ and $c = 4$ and as such is a false statement
- Solve via $(1 \quad 1)\begin{pmatrix} 1 \\ 1 \end{pmatrix}(c) = (1 \quad 1)\begin{pmatrix} 3 \\ 4 \end{pmatrix}$, so $2c = 7$ or $c = 3.5$
- Row scale the first equation by 10 to obtain: $\begin{pmatrix} 10 \\ 1 \end{pmatrix}(c) = \begin{pmatrix} 30 \\ 4 \end{pmatrix}$
- Solve via $(10 \quad 1)\begin{pmatrix} 10 \\ 1 \end{pmatrix}(c) = (10 \quad 1)\begin{pmatrix} 30 \\ 4 \end{pmatrix}$, so $101c = 304$ or $c = 3\frac{1}{101}$
- Perfectly valid row scaling leads to a different answer

False Statements

- Again, starting with the same: $\begin{pmatrix} 1 \\ 1 \end{pmatrix}(c) = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$
 - Subtract 2*(row 1) from row 2 to obtain $\begin{pmatrix} 1 \\ -1 \end{pmatrix}(c) = \begin{pmatrix} 3 \\ -2 \end{pmatrix}$
 - Solve via $(1 \quad -1) \begin{pmatrix} 1 \\ -1 \end{pmatrix}(c) = (1 \quad -1) \begin{pmatrix} 3 \\ -2 \end{pmatrix}$, so $2c = 5$ or $c = 2.5$
 - A perfectly valid row operation again leads to a different answer
 - Note that $2.5 \notin [3,4]$ either!
-
- Problem: $\begin{pmatrix} 3 \\ 4 \end{pmatrix}$ is not in the range of $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, so $\begin{pmatrix} 1 \\ 1 \end{pmatrix}(c) \neq \begin{pmatrix} 3 \\ 4 \end{pmatrix}$ for any $c \in \mathcal{R}$

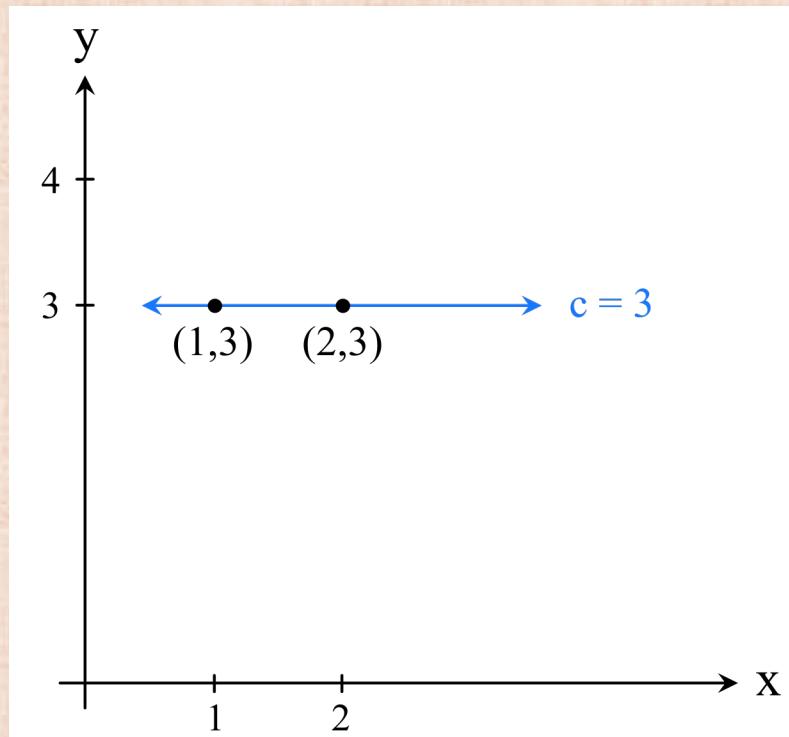
False Statements

- Consider $y = c_1 \phi_1$ with monomial $\phi_1 = 1$, and data points (1,3) and (2,4)
- This leads to the same $\begin{pmatrix} 1 \\ 1 \end{pmatrix} (c_1) = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$



True Statements

- Consider $y = c_1 \phi_1$ with monomial $\phi_1 = 1$, and data points $(1,3)$ and $(2,3)$
- This leads instead to $\binom{1}{1} (c_1) = \binom{3}{3}$ which is valid and has solution $c_1 = 3$



True Statements

- When b is in the range of A , then $Ac = b$ is a true statement
 - There exists at least one c (by definition) constrained by this statement
- When b is in not the range of A , then $Ac \neq b$ is the true statement
 - In this case, $Ac \neq b$ is true for all c
- The equation for the residual $r = b - Ac$ is always true
 - When b is in the range of A , there exists a c with $Ac = b$ and $r = 0$
 - When b is not in the range of A , then $Ac \neq b$ and $r \neq 0$ for all c
- The goal in both cases is to minimize the residual $r = b - Ac$

Norm Matters

- Consider $y = c_1 \phi_1$ where $\phi_1 = 1$ along with data points (1,3), (2,3), and (3,4)

- This leads to $r = \begin{pmatrix} 3 \\ 3 \\ 4 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}(c_1)$

- Setting $c_1 = 3.5$ minimizes $\|r\|_\infty$ with $r = \begin{pmatrix} -.5 \\ -.5 \\ .5 \end{pmatrix}$, $\|r\|_\infty = .5$, $\|r\|_2 = \frac{\sqrt{3}}{2}$

- Setting $c_1 = \frac{10}{3}$ minimizes $\|r\|_2$ with $r = \begin{pmatrix} -1/3 \\ -1/3 \\ 2/3 \end{pmatrix}$, $\|r\|_\infty = \frac{2}{3}$, $\|r\|_2 = \frac{\sqrt{6}}{3}$

Row Operations Matter

- Given a set of equations, they can be manipulated in various ways
- These manipulations may/typically change the answer
- Thus, one should carefully choose the residual they want minimized
- Equivalent sets of equations lead to different answers when minimizing the corresponding residuals

Weighted Minimization

- Given $r = b - Ac$, some equations may be deemed more important than others
- Scaling entries in the residual (before taking the norm) changes the relative importance of various equations
- I.e., minimize $\|Dr\|$ for a diagonal matrix D
- This is equivalent to row scaling: $Dr = Db - DAC$
- Column scaling doesn't effect the result, e.g. $Dr = Db - DAD^{-1}(\widehat{D}c)$
- So, it can be used to preserve symmetry: $Dr = Db - (DAD^T)(D^{-T}c)$

Least Squares

- Minimizing $\|r\|_2$ is referred to as least squares, and the resulting solution is referred to as the least squares solution
 - The least squares solution is the unique solution when $\|r\|_2 = 0$
- Minimizing $\|Dr\|_2$ is referred to as weighted least squares
- $\|r\|_2$ is minimized when $\|r\|_2^2$ is minimized
- And $\|r\|_2^2 = r \cdot r = (b - Ac) \cdot (b - Ac) = c^T A^T Ac - 2b^T Ac + b^T b$ is minimized when $c^T A^T Ac - 2b^T Ac$ is minimized
- Thus, minimize $c^T A^T Ac - 2b^T Ac$
- Similarly, for weighted least squares, minimize $c^T A^T D^2 Ac - 2b^T D^2 Ac$

Unit 9

Basic Optimization

Jacobian

- The Jacobian of $F(c) = \begin{pmatrix} F_1(c) \\ F_2(c) \\ \vdots \\ F_m(c) \end{pmatrix}$ has entries $J_{ik} = \frac{\partial F_i}{\partial c_k}(c)$

- Thus, the Jacobian $J(c) = F'(c) = \begin{pmatrix} \frac{\partial F_1}{\partial c_1}(c) & \frac{\partial F_1}{\partial c_2}(c) & \cdots & \frac{\partial F_1}{\partial c_n}(c) \\ \frac{\partial F_2}{\partial c_1}(c) & \frac{\partial F_2}{\partial c_2}(c) & \cdots & \frac{\partial F_2}{\partial c_n}(c) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial c_1}(c) & \frac{\partial F_m}{\partial c_2}(c) & \cdots & \frac{\partial F_m}{\partial c_n}(c) \end{pmatrix}$

Gradient

- Consider the scalar (output) function $f(c)$ with multi-dimensional input c
- The Jacobian of $f(c)$ is $J(c) = \begin{pmatrix} \frac{\partial f}{\partial c_1}(c) & \frac{\partial f}{\partial c_2}(c) & \cdots & \frac{\partial f}{\partial c_n}(c) \end{pmatrix}$
- The gradient of $f(c)$ is $\nabla f(c) = J^T(c) = \begin{pmatrix} \frac{\partial f}{\partial c_1}(c) \\ \frac{\partial f}{\partial c_2}(c) \\ \vdots \\ \frac{\partial f}{\partial c_n}(c) \end{pmatrix}$
- In 1D, both $J(c)$ and $\nabla f(c) = J^T(c)$ are the usual $f'(c)$

Critical Points

- To identify critical points of $f(c)$, set the gradient to zero: $\nabla f(c) = 0$

- This is a system of equations:

$$\begin{pmatrix} \frac{\partial f}{\partial c_1}(c) \\ \frac{\partial f}{\partial c_2}(c) \\ \vdots \\ \frac{\partial f}{\partial c_n}(c) \end{pmatrix} = 0 \quad \text{or}$$

$$\begin{pmatrix} \frac{\partial f}{\partial c_1}(c) = 0 \\ \frac{\partial f}{\partial c_2}(c) = 0 \\ \vdots \\ \frac{\partial f}{\partial c_n}(c) = 0 \end{pmatrix}$$

- Any c that simultaneously solves all the equations is a critical point
- In 1D, this is the usual $f'(c) = 0$

Hessian

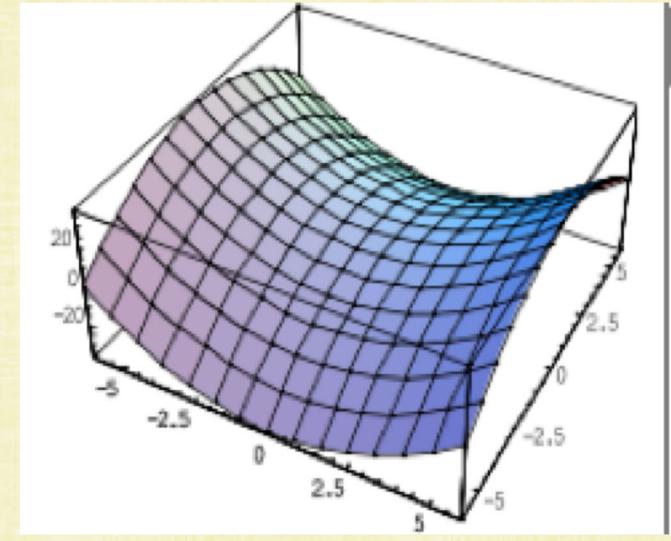
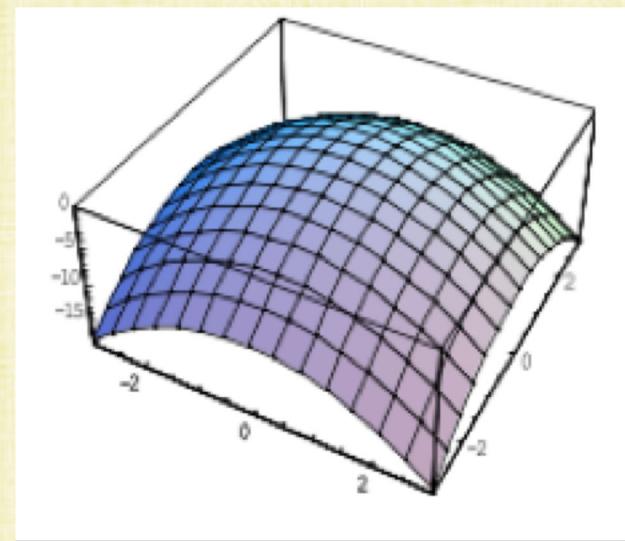
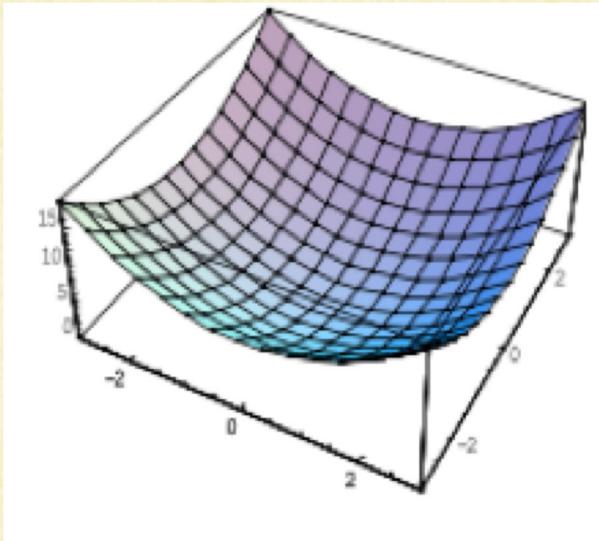
- The Hessian of $f(c)$ is $H(c) = J(\nabla f(c))^T$ and has entries $H_{ik} = \frac{\partial^2 f}{\partial c_i \partial c_k}(c)$
- The Hessian is $H(c) = \begin{pmatrix} \frac{\partial^2 f}{\partial c_1^2}(c) & \frac{\partial^2 f}{\partial c_1 \partial c_2}(c) & \cdots & \frac{\partial^2 f}{\partial c_1 \partial c_n}(c) \\ \frac{\partial^2 f}{\partial c_2 \partial c_1}(c) & \frac{\partial^2 f}{\partial c_2^2}(c) & \cdots & \frac{\partial^2 f}{\partial c_2 \partial c_n}(c) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial c_n \partial c_1}(c) & \frac{\partial^2 f}{\partial c_n \partial c_2}(c) & \cdots & \frac{\partial^2 f}{\partial c_n^2}(c) \end{pmatrix}$
- $H(c)$ is symmetric, when the order of differentiation doesn't matter
- In 1D, this is the usual $f''(c)$

Differential Forms

- Vector valued function: $dF(c) = J(F(c))dc$
- Scalar valued function (no different): $df(c) = J(f(c))dc$
- Transpose second equation (above): $df(c) = dc^T \nabla f(c)$
- Substitute ∇f for F (above): $d\nabla f(c) = J(\nabla f(c))dc$ or $d\nabla f(c) = H^T(c)dc$
- Take differential of $df(c)$ transposed: $d^2f(c) = J(dc^T \nabla f(c))dc$
- Some hand waving: $d^2f(c) = dc^T H^T(c)dc$

Classifying Critical Points

- Given critical point c^* , i.e. with $\nabla f(c^*) = 0$, the Hessian is used to classify it
- If $H(c^*)$ is positive definite, then c^* is a local minimum
- If $H(c^*)$ is negative definite, then c^* is a local maximum
- Otherwise, $H(c^*)$ is indefinite, and c^* is a saddle point



Classifying Critical Points (in 1D)

- In 1D, given critical point c^* , i.e. with $\nabla f(c^*) = f'(c^*) = 0$, the Hessian is used to classify it
- In 1D, $H(c^*) = (f''(c^*))$ is a size 1×1 diagonal matrix with eigenvalue $f''(c^*)$
- If $H(c^*)$ is positive definite with eigenvalue $f''(c^*) > 0$, then c^* is a local minimum
 - As usual, $f''(c^*) > 0$ implies concave up and a local min
- If $H(c^*)$ is negative definite with eigenvalue $f''(c^*) < 0$, then c^* is a local maximum
 - As usual, $f''(c^*) < 0$ implies concave down and a local max
- Otherwise, $H(c^*)$ is indefinite with eigenvalue $f''(c^*) = 0$, and c^* is a saddle point
 - As usual, $f''(c^*) = 0$ implies an inflection point (not a local extrema)

Quadratic Form

- The quadratic form of a square matrix \tilde{A} is $f(c) = \frac{1}{2}c^T \tilde{A}c - \tilde{b}^T c + \tilde{c}$
 - In 1D, $f(c) = \frac{1}{2}\tilde{a}c^2 - \tilde{b}c + \tilde{c}$
- Minimize $f(c)$ by (first) finding critical points where $\nabla f(c) = 0$
- Note $\nabla f(c) = \frac{1}{2}\tilde{A}c + \frac{1}{2}\tilde{A}^T c - \tilde{b}$, since $J(c^T v) = J(v^T c) = v^T$ (the gradient is v)
 - Solve the symmetric system $\frac{1}{2}(\tilde{A} + \tilde{A}^T)c = \tilde{b}$ to find critical points
- When \tilde{A} is symmetric, $\nabla f(c) = \tilde{A}c - \tilde{b} = 0$ is satisfied when $\tilde{A}c = \tilde{b}$
 - In 1D, the critical point is on the line of symmetry $\tilde{c} = \frac{\tilde{b}}{\tilde{a}}$
- That is, solve $\tilde{A}c = \tilde{b}$ to find the critical point

Quadratic Form

- The Hessian of $f(c)$ is $H = \frac{1}{2}(\tilde{A}^T + \tilde{A})$ or just \tilde{A} when \tilde{A} is symmetric
- When \tilde{A} is SPD, the solution to $\tilde{A}x = \tilde{b}$ is a minimum
- When \tilde{A} is symmetric negative definite, the solution to $\tilde{A}x = \tilde{b}$ is a maximum
- When \tilde{A} is indefinite, the solution to $\tilde{A}x = \tilde{b}$ is a saddle point
- In 1D, $H = (\tilde{a})$ is a size 1×1 diagonal matrix with eigenvalue \tilde{a}
- As usual, $\tilde{a} > 0$ implies concave up and a local min
- As usual, $\tilde{a} < 0$ implies concave down and a local max
- As usual, $\tilde{a} = 0$ implies an inflection point (not a local extrema)

Recall: Least Squares (Unit 8)

- Minimizing $\|r\|_2$ is referred to as least squares, and the resulting solution is referred to as the least squares solution
 - The least squares solution is the unique solution when $\|r\|_2 = 0$
- Minimizing $\|Dr\|_2$ is referred to as weighted least squares
- $\|r\|_2$ is minimized when $\|r\|_2^2$ is minimized
- And $\|r\|_2^2 = r \cdot r = (b - Ac) \cdot (b - Ac) = c^T A^T Ac - 2b^T Ac + b^T b$ is minimized when $c^T A^T Ac - 2b^T Ac$ is minimized
- Thus, minimize $c^T A^T Ac - 2b^T Ac$
- Similarly, for weighted least squares, minimize $c^T A^T D^2 Ac - 2b^T D^2 Ac$

Normal Equations

- $c^T A^T D^2 A c - 2b^T D^2 A c$ has the same minimum as $\frac{1}{2}c^T A^T D^2 A c - b^T D^2 A c$
- This is a quadratic form with symmetric $\tilde{A} = A^T D^2 A$ and $\tilde{b} = A^T D^2 b$
- The critical point is found from solving $\tilde{A}c = \tilde{b}$ or $A^T D^2 A c = A^T D^2 b$
- Weighted least squares defaults to ordinary least squares when $D = I$
- So, for (unweighted) least squares, solve $A^T A c = A^T b$
- These are called the normal equations

Hessian

- Recall: A is a tall (or square) full rank matrix with size $m \times n$ where $m \geq n$
- The Hessian $H = \tilde{A} = A^T A = V \Sigma^T U^T U \Sigma V^T = V \Sigma^T \Sigma V^T = V \Lambda V^T$
 - where $\Lambda = \Sigma^T \Sigma$ is a size $n \times n$ matrix of (nonzero) singular values squared
- $HV = V\Lambda$ illustrates that H has all positive eigenvalues (and so is SPD)
- That is, the critical point is indeed a minimum (as desired)
- For weighted least squares:
- Nonzero diagonal elements in D implies that $DAc = 0$ if and only if $Ac = 0$
- That is, a full column rank A implies a full column rank DA
- Then, the SVD of DA can be used to prove that $H = (DA)^T(DA)$ is SPD

Unit 10

Solving Least Squares

Normal Equations

- Let \tilde{A} have full column rank, and be size $m \times n$ with $m \geq n$
- Diagonal (nonzero) weighting $A = D\tilde{A}$ does not change the rank/size
 - but changes the answer when $D \neq I$ and $m \neq n$
- Minimizing $\|r\|_2 = \|b - Ac\|_2$ leads to the normal equations $A^T A c = A^T b$ for the critical point
- Since $A^T A$ is SPD, $A^T A c = A^T b$ has a unique solution obtainable via fast/efficient SPD solvers
- When b is in the range of A , the unique solution to $A^T A c = A^T b$ makes $r = 0$, and is thus the unique solution to $Ac = b$
 - When A is square ($m = n$), then b is always in the range of A

Condition Number

- Compare $A = U\Sigma V^T$ and $A^T A = V\Sigma^T \Sigma V^T = V\Lambda V^T$ where $\Lambda = \Sigma^T \Sigma$ is a diagonal size $n \times n$ matrix of singular values squared
- Since the singular values of $A^T A$ are the square of those in A , the condition number $\frac{\sigma_{max}}{\sigma_{min}}$ of $A^T A$ is also squared (compared to A)
 - And so, solving requires twice the machine precision (e.g. $(10^7)^2 = 10^{14}$)
- **It takes twice as much precision to get the same number of significant digits!**
- Thus, the normal equations should only be used as a last resort (when there are no other options)
- However, (like the SVD) it is a great tool for theoretical purposes
 - I.e. transform any full column rank matrix into an SPD system

Understanding Least Squares

- When $A = U\Sigma V^T$ has full (column) rank, $\Sigma = \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix}$ with $\hat{\Sigma}$ a size $n \times n$ diagonal matrix of positive singular values
 - The 0 term is size $(m - n) \times n$ and doesn't exist when $m = n$
- Then $A^T A = V \Sigma^T \Sigma V^T = V \hat{\Sigma}^2 V^T$ and $(A^T A)^{-1} = V \hat{\Sigma}^{-2} V^T$
- So $c = (A^T A)^{-1} A^T b = V \hat{\Sigma}^{-2} V^T V \Sigma^T U^T b = V (\hat{\Sigma}^{-1} \quad 0) U^T b$
- $Ac = U \Sigma V^T V (\hat{\Sigma}^{-1} \quad 0) U^T b = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} (\hat{\Sigma}^{-1} \quad 0) U^T b = U \begin{pmatrix} I_{n \times n} & 0 \\ 0 & 0 \end{pmatrix} U^T b$
- $r = b - Ac = UI_{m \times m} U^T b - U \begin{pmatrix} I_{n \times n} & 0 \\ 0 & 0 \end{pmatrix} U^T b = U \begin{pmatrix} 0 & 0 \\ 0 & I_{(m-n) \times (m-n)} \end{pmatrix} U^T b$

Understanding Least Squares

- From SVD slides (unit 3):
 - The columns of U corresponding to “nonzero” singular values form an orthonormal basis for the range of A
 - The remaining columns form an orthonormal basis for the space perpendicular to the range of A
- Since A only has n singular values, only the first n columns of U (which has m columns) span the range of A
- Writing $\begin{pmatrix} \hat{b}_r \\ \hat{b}_z \end{pmatrix} = U^T b$ where \hat{b}_r is size $n \times 1$ and \hat{b}_z is size $(m - n) \times 1$ separates out \hat{b}_r as the part of \hat{b} in the range of A
- Then (from prior slide): $c = V\hat{\Sigma}^{-1}\hat{b}_r$, $Ac = U\begin{pmatrix} \hat{b}_r \\ 0 \end{pmatrix}$, and $r = U\begin{pmatrix} 0 \\ \hat{b}_z \end{pmatrix}$

Orthogonal Matrices (and the L2 norm)

- From orthogonal matrices slides (unit 3):
 - Orthogonal matrices have orthonormal columns (an orthonormal basis), so their transpose is their inverse. They preserve inner products, and thus are rotations, reflections, and combinations thereof
- An orthogonal Q has $QQ^T = Q^TQ = I$
- So $\|Qr\|_2 = \sqrt{Qr \cdot Qr} = \sqrt{r^T Q^T Q r} = \sqrt{r^T r} = \|r\|_2$
- That is, orthogonal transformations preserve Euclidean distance

Understanding Least Squares

- $r = U \begin{pmatrix} 0 \\ \hat{b}_z \end{pmatrix}$ with orthogonal U implies $\|r\|_2 = \|\hat{b}_z\|_2$
- Consider the diagonal SVD view of $Ac = b$ when A has full rank
- That is, $U\Sigma V^T c = b$ or $\begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} \hat{c} = \begin{pmatrix} \hat{b}_r \\ \hat{b}_z \end{pmatrix}$
- The first block row gives $c = V\hat{\Sigma}^{-1}\hat{b}_r$, which is the least squares solution
- The second block row is $0 = \hat{b}_z$, and the norm of the residual for this block row is $\|\hat{b}_z\|_2$, which is identical to $\|r\|_2$
- The SVD approach gives the same (minimum residual) least squares solution

Recall: Gram-Schmidt

- Orthogonalizes a set of vectors
- For each vector, subtract its dot product overlap with all prior vectors, making it orthogonal to them
- A-orthogonal Gram-Schmidt simply uses A-weighted dot products
- Given vector \bar{S}^q , subtract out the A-overlap with s^1 to s^{q-1} so that the resulting vector s^q has $\langle s^q, s^{\hat{q}} \rangle_A = 0$ for $\hat{q} \in \{1, 2, \dots, q-1\}$
- That is, $s^q = \bar{S}^q - \sum_{\hat{q}=1}^{q-1} \frac{\langle \bar{S}^q, s^{\hat{q}} \rangle_A}{\langle s^{\hat{q}}, s^{\hat{q}} \rangle_A} s^{\hat{q}}$ where the two non-normalized $s^{\hat{q}}$ both require division by their norm (note that $\langle s^{\hat{q}}, s^{\hat{q}} \rangle_A = \|s^{\hat{q}}\|_A^2$)
- Proof: $\langle s^q, s^{\hat{q}} \rangle_A = \langle \bar{S}^q, s^{\hat{q}} \rangle_A - \frac{\langle \bar{S}^q, s^{\hat{q}} \rangle_A}{\langle s^{\hat{q}}, s^{\hat{q}} \rangle_A} \langle s^{\hat{q}}, s^{\hat{q}} \rangle_A = 0$

Gram-Schmidt QR Factorization

- From A , create a full rank Q with orthonormal columns
- For each column a_k , subtract the overlap with all prior columns in Q and make unit length:

$$q_k = \frac{a_k - \sum_{\hat{k}=1}^{k-1} \langle a_k, q_{\hat{k}} \rangle q_{\hat{k}}}{\|a_k - \sum_{\hat{k}=1}^{k-1} \langle a_k, q_{\hat{k}} \rangle q_{\hat{k}}\|_2}$$

- Define $r_{ik} = \langle a_i, q_k \rangle$ for $i > k$, and $r_{kk} = \|a_k - \sum_{\hat{k}=1}^{k-1} \langle a_k, q_{\hat{k}} \rangle q_{\hat{k}}\|_2$
- Then $q_k = \frac{a_k - \sum_{\hat{k}=1}^{k-1} r_{k\hat{k}} q_{\hat{k}}}{r_{kk}}$ and $a_k = r_{kk} q_k + \sum_{\hat{k}=1}^{k-1} r_{k\hat{k}} q_{\hat{k}} = \sum_{\hat{k}=1}^k r_{k\hat{k}} q_{\hat{k}}$
- That is $A = QR$ where R is upper triangular and $Q^T Q = I$

Gram-Schmidt QR (Example)

- Example: $A = QR$ with upper triangular R

$$\begin{pmatrix} 3 & -3 & 3 \\ 2 & -1 & 1 \\ 2 & -1 & -1 \\ 2 & -3 & 3 \\ 2 & -3 & 5 \end{pmatrix} = \begin{pmatrix} 3/5 & 0 & 0 \\ 2/5 & 1/2 & 1/2 \\ 2/5 & 1/2 & -1/2 \\ 2/5 & -1/2 & -1/2 \\ 2/5 & -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 5 & -5 & 5 \\ 0 & 2 & -4 \\ 0 & 0 & 2 \end{pmatrix}$$

- Note that $Q^T Q = I_{3 \times 3}$ since the columns of Q are orthonormal
- However, $QQ^T \neq I_{5 \times 5}$ and Q is not orthogonal

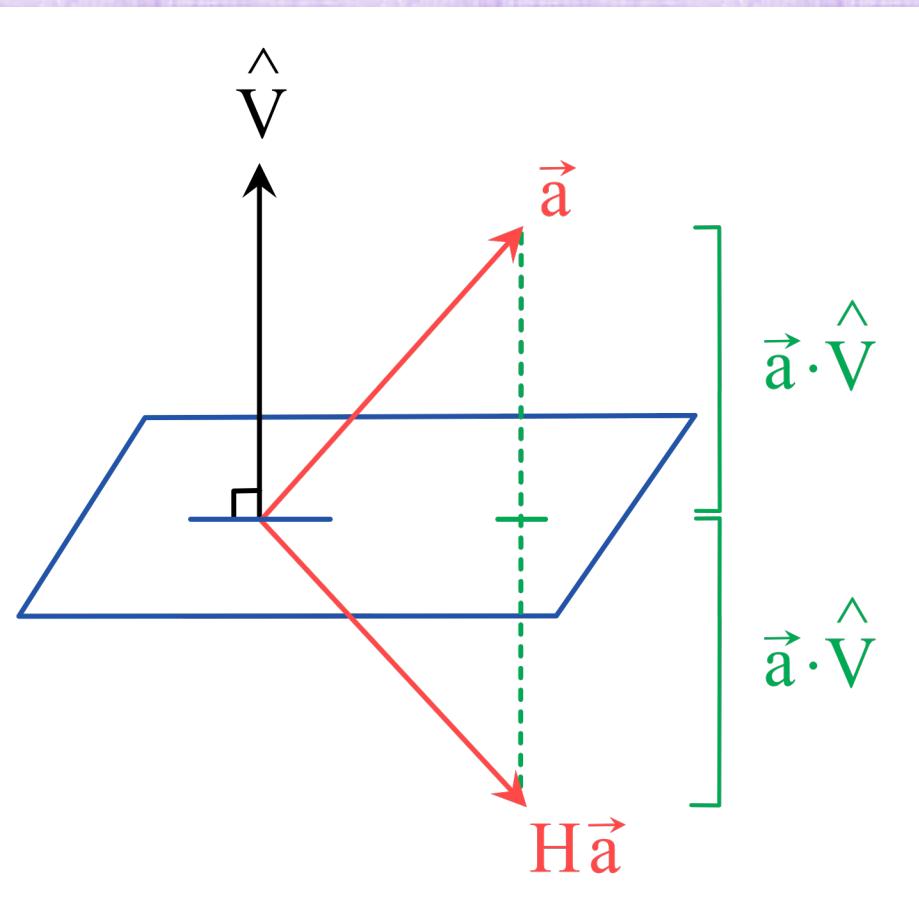
Not Good for Large Matrices

- Gram-Schmidt has too much numerical drift for large matrices
- Don't use Gram-Schmidt to find $A = QR$ with upper triangular R and $Q^T Q = I$
- But it does provide a good conceptual way to think about $A = QR$

QR Factorization

- Consider $A = QR$ with upper triangular R and $Q^T Q = I$
- Q is size $m \times n$ with n orthonormal columns
- Let \tilde{Q} be the matrix with $m - n$ orthonormal columns that span the space perpendicular to the range of Q
- Then the size $m \times m$ matrix $\hat{Q} = (Q \quad \tilde{Q})$ is orthogonal
- So $\|r\|_2 = \|\hat{Q}^T r\|_2 = \left\| \begin{pmatrix} Q^T \\ \tilde{Q}^T \end{pmatrix} (b - QRc) \right\|_2 = \left\| \begin{pmatrix} Q^T b - R c \\ \tilde{Q}^T b \end{pmatrix} \right\|_2$
- Since only the first block row varies, $\|r\|_2$ is minimized by solving $Rc = Q^T b$
- Since R is upper triangular, $Rc = Q^T b$ can be solved via back-substitution

Householder Transform



- Let unit normal \hat{v} implicitly define a plane orthogonal to it
- Then $H = I - 2\hat{v}\hat{v}^T$ reflects vectors across that plane
- $Ha = a - 2(\hat{v}^T a) \hat{v}$
- H is orthogonal with $H = H^T = H^{-1}$
- Usually don't form H , but (instead) only apply it via the definition of \hat{v}

Householder Transform

- Choose $v_k = a - Ha$ in a manner that zeroes out elements

$$\text{E.g. } v_k = \begin{pmatrix} a_1 \\ \vdots \\ a_{k-1} \\ a_k \\ a_{k+1} \\ \vdots \\ a_n \end{pmatrix} - \begin{pmatrix} a_1 \\ \vdots \\ a_{k-1} \\ \gamma \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \hat{a}_k - \gamma \hat{e}_k \text{ where } \hat{a}_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ a_k \\ a_{k+1} \\ \vdots \\ a_n \end{pmatrix}$$

- The vectors must have the same length (i.e. a reflection), so $\|\gamma\|_2 = \|\hat{a}_k\|_2$
- Thus, $v_k = \hat{a}_k \pm \|\hat{a}_k\|_2 \hat{e}_k$, which is subsequently normalized to $\hat{v}_k = \frac{v_k}{\|v_k\|_2}$
- For robustness, $v_k = \hat{a}_k + S(a_k) \|\hat{a}_k\|_2 \hat{e}_k$ where $S(a_k) = \pm 1$ is the sign function

Householder Transform (Example)

- Consider $a = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$ in the formulation $v_k = \hat{a}_k + S(a_k) \|\hat{a}_k\|_2 \hat{e}_k$
- Here $\hat{a}_1 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}$, $v_1 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} + S(2)\sqrt{9}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix}$, $\hat{v}_1 = \frac{1}{\sqrt{30}}\begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix}$
- So $H_1 a = a - 2(\hat{v}_1^T a) \hat{v}_1 = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix} - 2 \frac{15}{\sqrt{30}} \frac{1}{\sqrt{30}} \begin{pmatrix} 5 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} -3 \\ 0 \\ 0 \end{pmatrix}$

Householder Transform (QR)

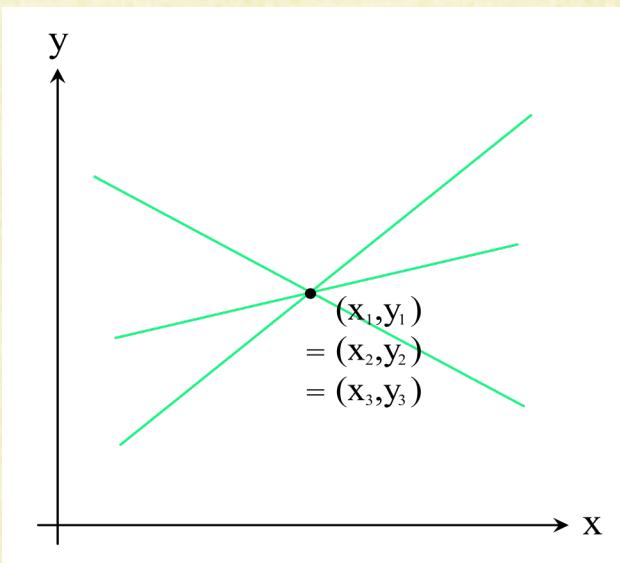
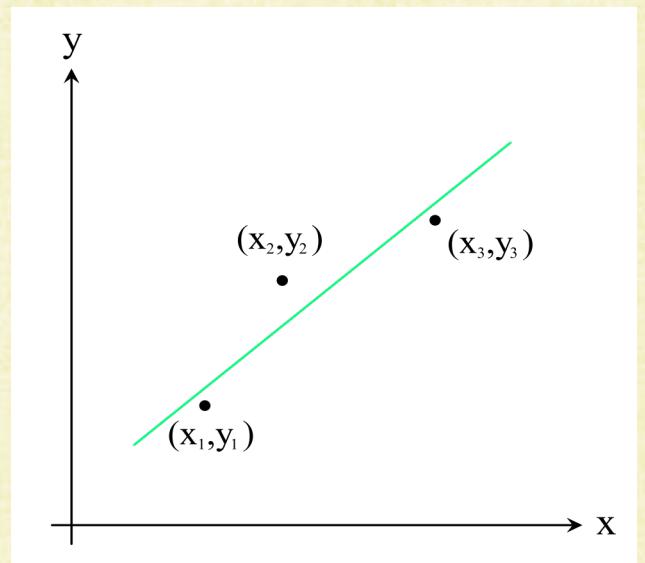
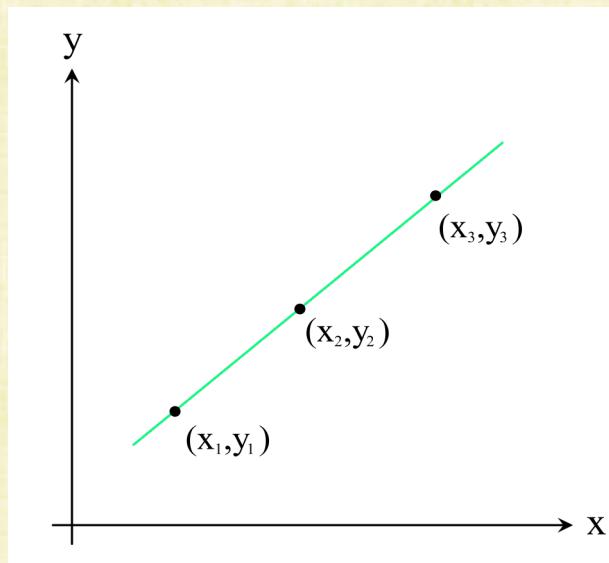
- For each column of A , construct the Householder transform that zeroes out entries below the diagonal
- Then $H_n H_{n-1} \cdots H_2 H_1 A = \begin{pmatrix} R \\ 0 \end{pmatrix}$ and $H_n H_{n-1} \cdots H_2 H_1 b = \begin{pmatrix} \tilde{b}_r \\ \tilde{b}_z \end{pmatrix}$
- Apply H_k efficiently using \hat{v}_k and to apply it to all columns $\geq k$
 - It doesn't affect columns $< k$
- Note that H_n is required to get zeroes at the bottom of the last column
- Letting $Q^T = H_n H_{n-1} \cdots H_2 H_1$ gives $Q^T A = \begin{pmatrix} R \\ 0 \end{pmatrix}$ or $A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$
- And $\|r\|_2 = \|Q^T r\|_2 = \left\| Q^T \left(b - Q \begin{pmatrix} R \\ 0 \end{pmatrix} c \right) \right\|_2 = \left\| \begin{pmatrix} \tilde{b}_r \\ \tilde{b}_z \end{pmatrix} - \begin{pmatrix} Rc \\ 0 \end{pmatrix} \right\|_2$
- So, minimize $\|r\|_2$ to $\|\tilde{b}_z\|_2$ by solving $Rc = \tilde{b}_r$

Unit 11

Zero Singular Values

Underdetermined Systems

- Consider drawing a line $y = c_1 + c_2x$ through 3 points
- When the points are colinear, there is a unique solution
- When the points are not colinear, there is a least squares solution
- When the points are co-located (i.e. identical), there are infinite solutions



Underdetermined Systems

- The Vandermonde matrix equation is $\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$
- When $x_1 = x_2 = x_3$, the columns are multiples of each other (matrix is rank 1)
- When $y_1 = y_2 = y_3$, the right hand side is in the range of the column space, and there are infinite solutions
- Otherwise, the right hand side is not in the range of the column space, and there are no solutions

Misleading Labels

- Consider $\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 0 \end{pmatrix}$
- The first two rows, $c_1 = 1$ and $c_1 = 2$, overdetermine c_1
- The third row, $c_2 = 3$, uniquely determines c_2
- The last row (column), $0c_3 = 0$, leaves c_3 underdetermined with infinite possibilities
- Misleading to classifying an entire system as either unique solution, no solution, or infinite solutions
- Rather, do the best one can with what has been given
 - E.g. Shouldn't skip dinner because of uncertainties about when the sun goes down

SVD (general framework)

- Transform $Ac = b$ into $\Sigma\hat{c} = \hat{b}$ (as usual)
- For each $\sigma_k \neq 0$, compute $\hat{c}_k = \frac{\hat{b}_k}{\sigma_k}$ (as usual)
- When $\sigma_k = 0$, \hat{c}_k is undefined (moreover, division by a small σ_k is dubious)
- A tall matrix has extra rows with $0 = \hat{b}_k$, and nonzero \hat{b}_k implies a nonzero residual
- A wide matrix has extra columns of zeros that leave some \hat{c}_k undetermined (no different than a $\sigma_k = 0$ column)

SVD (general framework)

- For tall matrices $A = U \begin{pmatrix} \hat{\Sigma} \\ 0 \end{pmatrix} V^T$, and for wide matrices $A = U(\hat{\Sigma} \quad 0)V^T$
- In both cases, $\hat{\Sigma}$ may contain zeros on the diagonal
- In general, can write $A = U \begin{pmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} V^T$ with $\hat{\Sigma}$ diagonal and full rank
- Then $\Sigma \hat{c} = \hat{b}$ has the form $\begin{pmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{c}_r \\ \hat{c}_z \end{pmatrix} = \begin{pmatrix} \hat{b}_r \\ \hat{b}_z \end{pmatrix}$
- $\|r\|_2 = \|U^T(b - Ac)\|_2 = \left\| \begin{pmatrix} \hat{b}_r \\ \hat{b}_z \end{pmatrix} - \begin{pmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{c}_r \\ \hat{c}_z \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} \hat{b}_r \\ \hat{b}_z \end{pmatrix} - \begin{pmatrix} \hat{\Sigma} \hat{c}_r \\ 0 \end{pmatrix} \right\|_2$
- So \hat{c}_r determined via $\hat{\Sigma} \hat{c}_r = \hat{b}_r$ minimizes the residual to $\|r\|_2 = \|\hat{b}_z\|_2$
- Meanwhile, any values are acceptable for \hat{c}_z

Minimum Norm Solution

- Since any values are acceptable for \hat{c}_z , set $\hat{c}_z = 0$ in order to indicate that these parameters have no bearing on the solution
- This is more sensible than setting \hat{c}_z to some nonzero value as if those values mattered
- Example:
 - Consider a variable related to how a hat is worn while driving, which could matter when the hat blocks the sun or keeps hair away from the eyes
 - Someone with short hair driving at night would likely have no driving dependence on the hat; in this case, reporting information on the hat is unimportant/misleading
- So, $c = V\hat{c} = V \begin{pmatrix} \hat{c}_r \\ \hat{c}_z \end{pmatrix} = V \begin{pmatrix} \hat{\Sigma}^{-1} \hat{b}_r \\ 0 \end{pmatrix} = \sum_{\sigma_k \neq 0} \nu_k \frac{\hat{b}_k}{\sigma_k} = \sum_{\sigma_k \neq 0} \nu_k \frac{u_k^T b}{\sigma_k}$

Pseudo-Inverse

- From the minimum norm solution, $c = \left(\sum_{\sigma_k \neq 0} \frac{v_k u_k^T}{\sigma_k} \right) b$
- That is, $c = A^+ b$ with pseudo-inverse $A^+ = \sum_{\sigma_k \neq 0} \frac{v_k u_k^T}{\sigma_k}$
- When A is square and full rank $A^{-1} = A^+$
- Each term is an outer product between corresponding columns of U and V weighted by one over their corresponding singular value
- Each term is a matrix of size $n \times m$, so this a sum of matrices

Sum of Rank One Matrices

- $Ac = U \begin{pmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} V^T c = U \begin{pmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{c}_r \\ \hat{c}_z \end{pmatrix} = U \begin{pmatrix} \hat{\Sigma} \hat{c}_r \\ 0 \end{pmatrix} = \sum_{\sigma_k \neq 0} u_k \sigma_k \hat{c}_k = \sum_{\sigma_k \neq 0} u_k \sigma_k v_k^T c = (\sum_{\sigma_k \neq 0} u_k \sigma_k v_k^T) c$
- Thus, $A = \sum_{\sigma_k \neq 0} \sigma_k u_k v_k^T$
- Each term is an outer product between corresponding columns of U and V weighted by their corresponding singular value
- Each term is a matrix of size $m \times n$ (the same size as A)
- Each term is rank 1, since every column in the matrix is a multiple of u_k

Recall: Understanding Ac (unit 3)

$$\begin{aligned}
 Ac &= \begin{pmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{pmatrix} \begin{pmatrix} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} .504 & .574 & .644 \\ -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} \\
 &= \begin{pmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{pmatrix} \begin{pmatrix} 25.5 & 0 & 0 \\ 0 & 1.29 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} v_1^T c \\ v_2^T c \\ v_3^T c \end{pmatrix} \\
 &= \begin{pmatrix} .141 & .825 & -.420 & -.351 \\ .344 & .426 & .298 & .782 \\ .547 & .028 & .644 & -.509 \\ .750 & -.371 & -.542 & .079 \end{pmatrix} \begin{pmatrix} \sigma_1 v_1^T c \\ \sigma_2 v_2^T c \\ \sigma_3 v_3^T c \\ 0 \end{pmatrix} \\
 &= u_1 \sigma_1 v_1^T c + u_2 \sigma_2 v_2^T c + u_3 \sigma_3 v_3^T c + u_4 0
 \end{aligned}$$

- Ac projects c onto the basis vectors in V , then scales by the associated singular values, and lastly uses those results as weights on the basis vectors in U

Matrix Approximation

- Use the p largest singular values: $A \approx \sum_{k=1}^p \sigma_k u_k v_k^T$
- The pseudo-inverse is approximated similarly: $A^+ \approx \sum_{k=1}^p \frac{v_k u_k^T}{\sigma_k}$
- This is the best rank p approximation to A , and the main idea behind principle component analysis (PCA)
 - Often, thousands/millions of terms can be thrown away keeping only 10 to 100 terms
- Drop small singular values: $A \approx \sum_{\sigma_k > \epsilon} \sigma_k u_k v_k^T$
- This makes the pseudo-inverse better conditioned: $A^+ \approx \sum_{\sigma_k > \epsilon} \frac{v_k u_k^T}{\sigma_k}$
 - Of course, this relies on a good choice of $\epsilon > 0$

Recall: Approximating A (unit 3)

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \end{pmatrix} \approx \left(\begin{array}{c|cc|cc} .141 & .825 & -.420 & -.351 & \\ \hline .344 & .426 & .298 & .782 & \\ .547 & .028 & .644 & -.509 & \\ .750 & -.371 & -.542 & .079 & \end{array} \right) \left(\begin{array}{cc|c} 25.5 & 0 & b \\ 0 & 1.29 & b \\ 0 & 0 & b \\ 0 & 0 & b \end{array} \right) \left(\begin{array}{ccc} .504 & .574 & .644 \\ \hline -.761 & -.057 & .646 \\ .408 & -.816 & .408 \end{array} \right)$$

- The first singular value is much bigger than the second, and so represents the vast majority of what A does (note, the vectors in U and V are unit length)
- Thus, one could approximate A quite well by only using the terms associated with the largest singular value
- This is not a valid factorization, but an approximation (and the idea behind PCA)

Rank One Updates

- For real time applications (e.g. real time decision making), iteratively add one term at a time (slowly improving the estimate)
- $c = A^+b \approx \frac{u_1^T b}{\sigma_1} v_1 + \frac{u_2^T b}{\sigma_2} v_2 + \frac{u_3^T b}{\sigma_3} v_3 + \dots$
- Note the efficient order of operations:
 - $u_k^T b$ is m multiplies and the result times v_k is n multiplies for a total of $m + n$
 - Instead, multiplying the size mxn matrix $v_k u_k^T$ times b is $m \cdot n$ multiplies

Computing the SVD

- $A^T A = V \Sigma^T \Sigma V^T$ so $(A^T A)V = V(\Sigma^T \Sigma)$
- $AA^T = U \Sigma \Sigma^T U^T$ so $(AA^T)U = U(\Sigma \Sigma^T)$
- If $\sigma_k \neq 0$, then σ_k^2 is an eigenvalue of both $A^T A$ and AA^T (with eigenvectors v_k and u_k respectively)
- Easier to work with the smaller of $A^T A$ and AA^T (which are both SP(S)D) to find each eigenvalue σ_k^2
- Then σ_k^2 can be used in both $A^T A$ and AA^T to find the corresponding eigenvectors

Condition Number of Eigenproblems

- The condition number for finding an eigenvalue is different than the condition number for solving a linear system
- The condition number for finding an eigenvalue/eigenvector pair is $\frac{1}{v_L^T v_R}$ where v_L and v_R are the normalized left/right eigenvectors
- Symmetric (Hermitian) matrices have left/right eigenvectors that are identical, so $v_L^T v_R = 1$ and the condition number is 1

Characteristic Polynomial

- The eigenvalue problem is typically written as $A\nu = \lambda\nu$
- Alternatively, $(A - \lambda I)\nu = 0$
- This is true when $\det(A - \lambda I) = 0$, which leads to a degree n characteristic polynomial equation in λ (for a size $n \times n$ matrix A)
- Finding the roots of this equation can be quite difficult
 - Recall how difficult it was to find roots for a mere cubic equation
- Finding roots for $n > 3$ is undesirable

Similarity Transforms

- Similarity transforms, $T^{-1}AT$, preserve the eigenstructure
 - $T^{-1}ATv = \lambda v$ or $A(Tv) = \lambda(Tv)$ still has eigenvalue λ with modified eigenvector Tv
- When A is real/symmetric (complex/Hermitian), an orthogonal (unitary) T exists to make $T^{-1}AT$ diagonal with real eigenvalues
 - e.g. $T = V$ for $A^T A = V \Sigma^T \Sigma V^T$ and $T = U$ for $AA^T = U \Sigma \Sigma^T U^T$
- When A has distinct eigenvalues, a T exists to make $T^{-1}AT$ diagonal
- For any matrix, a unitary T exists to make $T^{-1}AT$ upper triangular (Schur form) which has eigenvalues on the diagonal
- Any matrix can be put into Jordan form where the eigenvalues are on the diagonal, and off diagonal elements only occur on the band above the diagonal and only for defective eigenvalues (repeated eigenvalues that don't possess a full set of eigenvectors)

QR Iteration (for eigenvalues)

- Starting with $A^0 = A$
- Compute the factorization $A^q = Q^q R^q$
- Then define $A^{q+1} = R^q Q^q$
- Note that $R^q Q^q = (Q^q)^T Q^q R^q Q^q = (Q^q)^T A^q Q^q$ is a similarity transform of A^q
- If the eigenvalues are distinct, A^q converges to a triangular matrix
- If A is symmetric, A^q converges to a diagonal matrix

Power Method

- Computes the largest eigenvalue/eigenvector (great for rank 1 updates)
- Starting from $c^0 \neq 0$, iterate $c^{q+1} = Ac^q$
- Suppose c^0 is a linear combination of eigenvectors: $c^0 = \sum_k \alpha_k v_k$
- Then $c^q = A^q c^0 = \sum_k \alpha_k A^q v_k = \sum_k \alpha_k \lambda_k^q v_k = \lambda_{max}^q \sum_k \alpha_k \left(\frac{\lambda_k}{\lambda_{max}}\right)^q v_k$
- As $q \rightarrow \infty$, $\left(\frac{\lambda_k}{\lambda_{max}}\right)^q \rightarrow 0$ for $\lambda_k < \lambda_{max}$ and $c^q \rightarrow \lambda_{max}^q \alpha_{max} v_{max}$
- Thus, as $q \rightarrow \infty$, $\frac{(c^{q+1})_i}{(c^q)_i} \rightarrow \lambda_{max}$ for each of the i components of c
- Deflation removes an eigenvalue from A by subtracting off its rank 1 update
 - The deflated $A - (\sigma_k u_k v_k^T)$ can be used to compute the next largest eigenvalue, etc.

Power Method

- If $c^0 = \sum_k \alpha_k v_k$ happens to have $\alpha_{max} = 0$ the method might fail (but roundoff errors can help)
- When c^0 and A are real valued, cannot obtain complex numbers
- When the largest eigenvalue is repeated, the final vector may be a linear combination of the multiple eigenvectors
- c^q needs to be periodically renormalized to stop it from growing too large
- Inverse Iteration can be used to find the smallest eigenvalue of A , since the largest eigenvalue of A^{-1} is the smallest eigenvalue of A
 - Useful for finding the condition number $\frac{\sigma_{max}}{\sigma_{min}}$

Unit 12

Regularization

Adding the Identity

- Add $Ic = 0$ to drive components related to small/zero singular values to zero
 - Motivated by minimal norm solution
- Combine with the original system $\begin{pmatrix} A \\ I \end{pmatrix} c = \begin{pmatrix} b \\ 0 \end{pmatrix}$ so that $\begin{pmatrix} A \\ I \end{pmatrix}$ has full column rank
 - Can be solved with Householder, etc.
- The normal equations are $(A^T \quad I) \begin{pmatrix} A \\ I \end{pmatrix} c = (A^T \quad I) \begin{pmatrix} b \\ 0 \end{pmatrix}$ or $(A^T A + I)c = A^T b$
- Using $A = U\Sigma V^T$, this is $(V\Sigma^T \Sigma V^T + I)c = V\Sigma^T \hat{b}$ or $(\Sigma^T \Sigma + I)\hat{c} = \Sigma^T \hat{b}$
- The augmented $\sigma_k^2 + 1$ drive components with $\sigma_k = 0$ to zero (as desired)
- However, nonzero σ_k have their (unique or least squares) solution perturbed as well (since $Ic = 0$ interferes with $Ac = b$)

Perturbation

- Recall the most general $\Sigma = \begin{pmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{pmatrix}$ with $\hat{\Sigma}$ diagonal and full rank
- $(\Sigma^T \Sigma + I) \hat{c} = \Sigma^T \hat{b}$ becomes $\left(\begin{pmatrix} \hat{\Sigma}^T & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{pmatrix} + I \right) \begin{pmatrix} \hat{c}_r \\ \hat{c}_z \end{pmatrix} = \begin{pmatrix} \hat{\Sigma}^T & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{b}_r \\ \hat{b}_z \end{pmatrix}$
- Or $\left(\begin{pmatrix} \hat{\Sigma}^2 & 0 \\ 0 & 0 \end{pmatrix} + I \right) \begin{pmatrix} \hat{c}_r \\ \hat{c}_z \end{pmatrix} = \begin{pmatrix} \hat{\Sigma} \hat{b}_r \\ 0 \end{pmatrix}$ where $\hat{c}_z = 0$ as desired
- However, for the \hat{c}_r terms, $\hat{c}_k = \frac{\sigma_k}{\sigma_k^2 + 1} \hat{b}_k$ instead of $\hat{c}_k = \frac{1}{\sigma_k} \hat{b}_k$
- This perturbs \hat{c}_k away from its actual (unique or least squares) solution

Regularization

- For larger $\sigma_k \gg 1$, $\frac{\sigma_k}{\sigma_k^2 + 1} \approx \frac{1}{\sigma_k}$ and the perturbation of the (unique or least squares) solution is negligible
- But for $\sigma_k \approx 1$, the perturbation is quite large
- And for $\sigma_k \ll 1$, $\frac{\sigma_k}{\sigma_k^2 + 1} \approx 0$ drives \hat{c}_k towards zero
- Adding $\epsilon I c = 0$ (with $\epsilon > 0$) instead of $I c = 0$ leads to $\hat{c}_k = \frac{\sigma_k}{\sigma_k^2 + \epsilon^2} \hat{b}_k$
- This has limited effect on $\sigma_k \gg \epsilon$, but helps stabilize/regularize the solution for $\sigma_k < \epsilon$

Initial Guess

- $Ic = 0$ implies an initial guess of $c = 0$
- Instead suppose one had an initial guess of $c = c^*$
- Add $Ic = c^*$ to the equations (instead of $Ic = 0$), i.e. $\begin{pmatrix} A \\ I \end{pmatrix} c = \begin{pmatrix} b \\ c^* \end{pmatrix}$
- Normal equations are $(A^T A + I)c = A^T b + c^*$
- This leads to $(\Sigma^T \Sigma + I)\hat{c} = \Sigma^T \hat{b} + V^T c^* = \Sigma^T \hat{b} + \hat{c}^*$
- This gives $\hat{c}_k = \frac{\sigma_k}{\sigma_k^2 + 1} \hat{b}_k + \frac{1}{\sigma_k^2 + 1} \hat{c}_k^*$ which tends towards \hat{b}_k for larger σ_k as desired but tends towards \hat{c}_k^* for smaller σ_k (and $\hat{c}_k = \hat{c}_k^*$ when $\sigma_k = 0$)
- Adding $\epsilon Ic = \epsilon c^*$ gives $\hat{c}_k = \frac{\sigma_k}{\sigma_k^2 + \epsilon^2} \hat{b}_k + \frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \hat{c}_k^*$

Initial Guess

- Rewrite this as $\hat{c}_k = \left(\frac{\sigma_k^2}{\sigma_k^2 + \epsilon^2} \right) \frac{\hat{b}_k}{\sigma_k} + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right) \hat{c}_k^*$
 - Note the convex weights: $\left(\frac{\sigma_k^2}{\sigma_k^2 + \epsilon^2} \right) + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right) = 1$
- This is a convex combination of the usual solution $\frac{\hat{b}_k}{\sigma_k}$ and the initial guess \hat{c}_k^*
 - Also valid for initial guess of $\hat{c}_k^* = 0$
- Large σ_k as compared to ϵ tend toward the usual solution: $\hat{c}_k = \frac{\hat{b}_k}{\sigma_k}$
- Small σ_k as compared to ϵ tend toward the initial guess: $\hat{c}_k = \hat{c}_k^*$

Iterate

- First, solve with $\epsilon Ic = 0$ to get $\hat{c}_k = \left(\frac{\sigma_k^2}{\sigma_k^2 + \epsilon^2} \right) \frac{\hat{b}_k}{\sigma_k}$
- Then, use this solution as an initial guess and solve again to get:

$$\hat{c}_k = \left(\frac{\sigma_k^2}{\sigma_k^2 + \epsilon^2} \right) \frac{\hat{b}_k}{\sigma_k} + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right) \left(\frac{\sigma_k^2}{\sigma_k^2 + \epsilon^2} \right) \frac{\hat{b}_k}{\sigma_k} = \left(1 + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right) \right) \left(\frac{\sigma_k^2}{\sigma_k^2 + \epsilon^2} \right) \frac{\hat{b}_k}{\sigma_k}$$

- Then, use this solution as an initial guess and solve again to get:

$$\begin{aligned}\hat{c}_k &= \left(\frac{\sigma_k^2}{\sigma_k^2 + \epsilon^2} \right) \frac{\hat{b}_k}{\sigma_k} + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right) \left(1 + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right) \right) \left(\frac{\sigma_k^2}{\sigma_k^2 + \epsilon^2} \right) \frac{\hat{b}_k}{\sigma_k} \\ &= \left(1 + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right) + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right)^2 \right) \left(\frac{\sigma_k^2}{\sigma_k^2 + \epsilon^2} \right) \frac{\hat{b}_k}{\sigma_k}\end{aligned}$$

Continue Iterating

- Continuing leads to $\hat{c}_k = \left(1 + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right) + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right)^2 + \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2} \right)^3 + \dots \right) \left(\frac{\sigma_k^2}{\sigma_k^2 + \epsilon^2} \right) \frac{\hat{b}_k}{\sigma_k}$
- The geometric series in parenthesis has $r = \frac{\epsilon^2}{\sigma_k^2 + \epsilon^2}$
- It converges to $\frac{1}{1-r} = \frac{\sigma_k^2 + \epsilon^2}{\sigma_k^2}$ giving $\hat{c}_k = \frac{\hat{b}_k}{\sigma_k}$ in the limit (as desired)
- When $\sigma_k = 0$, the convex weights are 0 and 1, so $\hat{c}_k = 0$ identically at every step
 - Minimum norm solution for these σ_k
- As $\sigma_k \rightarrow 0$, $\hat{c}_k \rightarrow (1 + 1 + \dots + 1) \left(\frac{\sigma_k}{\epsilon^2} \right) \hat{b}_k \rightarrow 0$ for a finite number of steps
 - Small σ_k are regularized, sending their associated $\hat{c}_k \rightarrow 0$

Convergence Rate

- After q iterations, the geometric series sum is $\frac{1-r^q}{1-r} = \frac{\sigma_k^2 + \epsilon^2}{\sigma_k^2} \left(1 - \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2}\right)^q\right)$
- This gives $\hat{c}_k = \left(1 - \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2}\right)^q\right) \frac{\hat{b}_k}{\sigma_k}$ implying monotonic convergence to $\hat{c}_k = \frac{\hat{b}_k}{\sigma_k}$
 - Because $r = \left(\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2}\right) < 1$ implies $r^q \rightarrow 0$ as $q \rightarrow \infty$, and thus $(1 - r^q) \rightarrow 1$
- The convergence is quick for large σ_k (as desired)
- Smaller σ_k have $\frac{\epsilon^2}{\sigma_k^2 + \epsilon^2}$ closer to 1, so that their \hat{c}_k starts near zero and more slowly increases towards $\frac{\hat{b}_k}{\sigma_k}$
 - Smaller σ_k are regularized in this way

Comparison with PCA

- After q iterations, PCA includes the q largest σ_k components into the solution
- But PCA does not include any contribution for components with $k > q$
 - Larger components are all Heaviside thresholded to be identically zero
- After q iterations, this iterative method does not include the full contribution of the q largest σ_k components into the solution
 - It only includes $(1 - r_k^q)$ times those components, but $(1 - r_k^q)$ is close to 1 when σ_k is larger
- The iterative method includes contributions for components with $k > q$ too
 - Their contribution is smaller since their $(1 - r_k^q)$ is not as close to 1 when σ_k is smaller
 - The iterative method has a smoother falloff for increasing σ_k

Aside: Class Bonus

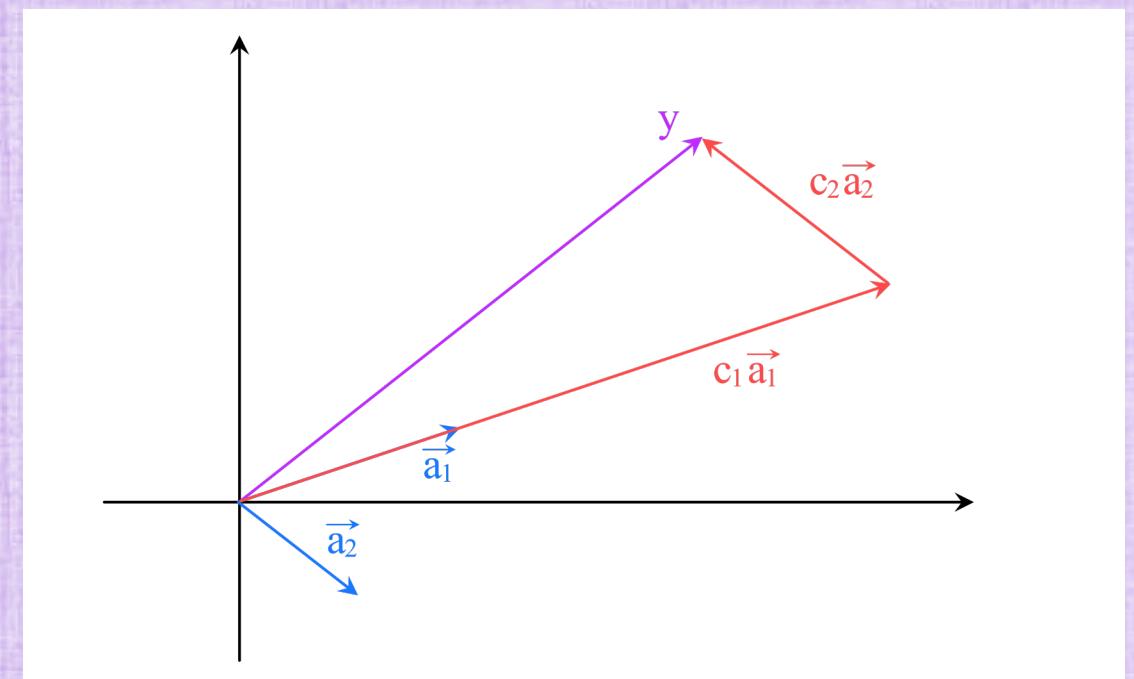
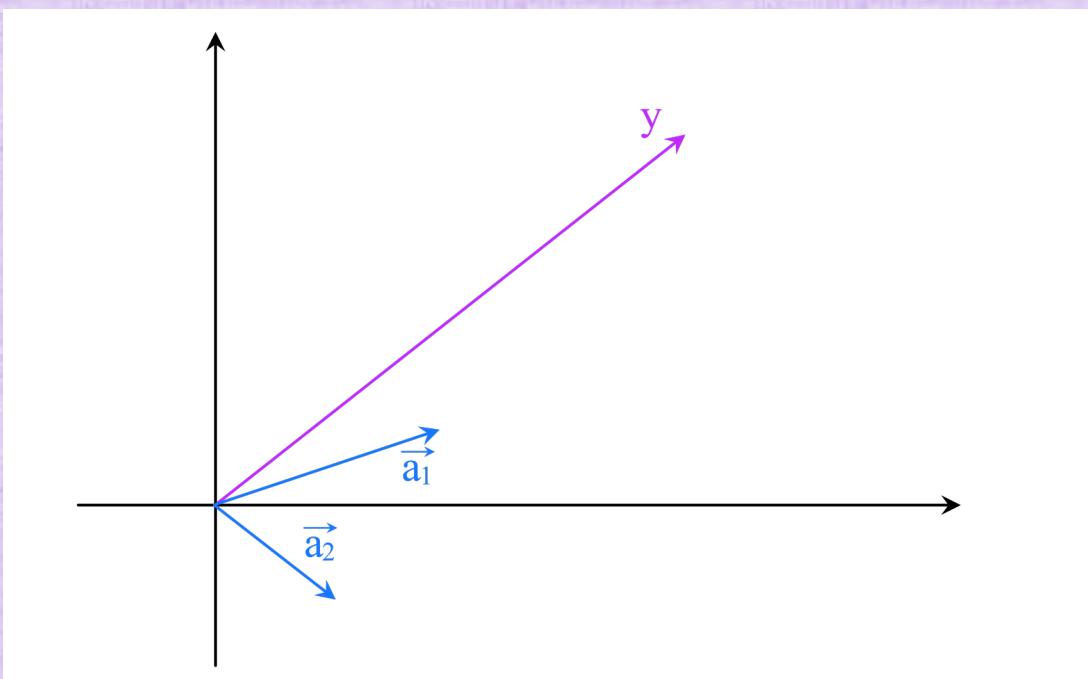
- This iterative method and the analysis via a geometric series (slides 7-10) were derived in preparation for this course last year (by me, for fun)
- In the recent past, Dan Boneh had complained (to me) that too much Calculus was being taught
 - and that, for example, even some CS professors didn't know the difference between algebraic/geometric series (off the top of their head)
- Interestingly, machine learning has changed this perception of "too much calculus", since calculus at the core of how we train neural networks
- Bounty: find this iterative method (and analysis) in the literature, and we will add 10 points (on a scale of 0-100) to your grade
 - The non-iterative version of the method is basically Levenberg-Marquardt

Adding a Diagonal Matrix

- Adding $Dc = 0$ to obtain: $\begin{pmatrix} A \\ D \end{pmatrix} c = \begin{pmatrix} b \\ 0 \end{pmatrix}$ weights some variables more strongly towards zero than others
- The normal equations are $(A^T A + D^2)c = A^T b$
- Equivalently $(V\Sigma^T \Sigma V^T + D^2)c = V\Sigma^T \hat{b}$ or $(\Sigma^T \Sigma + V^T D^2 V)\hat{c} = \Sigma^T \hat{b}$
- This last version of the normal equations also results from $\begin{pmatrix} \Sigma \\ DV \end{pmatrix} \hat{c} = \begin{pmatrix} \hat{b} \\ 0 \end{pmatrix}$
 - Unfortunately, D shears the vectors in V creating issues
- This motivates column scaling the original equations $\begin{pmatrix} AD^{-1} \\ I \end{pmatrix} Dc = \begin{pmatrix} b \\ 0 \end{pmatrix}$ so that the resulting $\begin{pmatrix} \tilde{A} \\ I \end{pmatrix} \tilde{c} = \begin{pmatrix} b \\ 0 \end{pmatrix}$ can be treated in the usual way with $I\tilde{c} = 0$

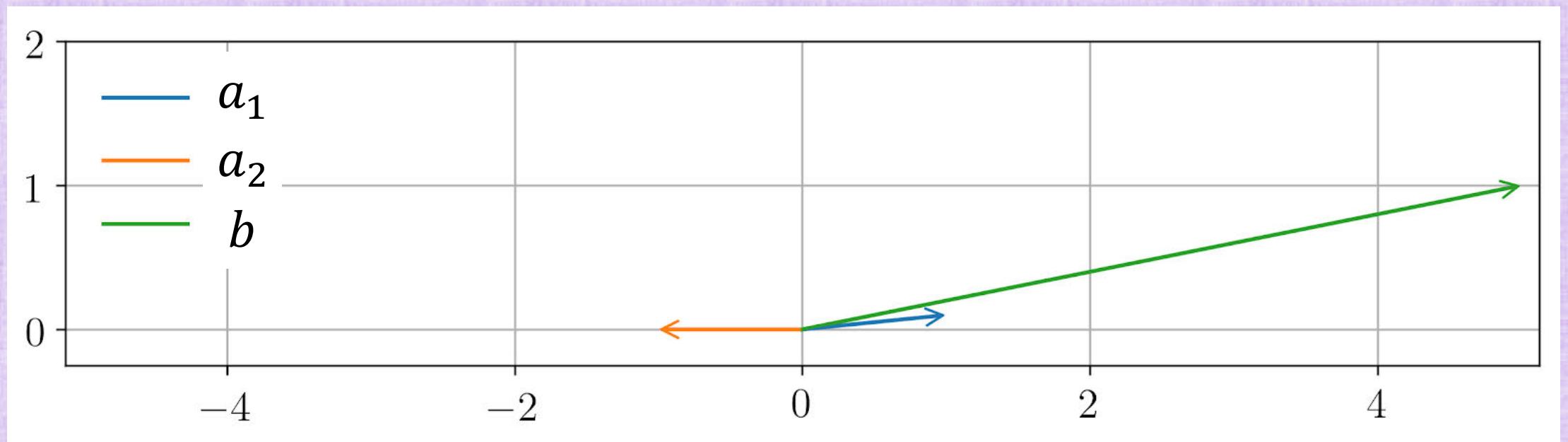
Recall: Matrix Columns as Vectors (unit 1)

- Let the k -th column of A be vector a_k , so $Ac = y$ is equivalent to $\sum_k c_k a_k = y$
- That is, find a linear combination of the columns of A that gives the right hand side vector y



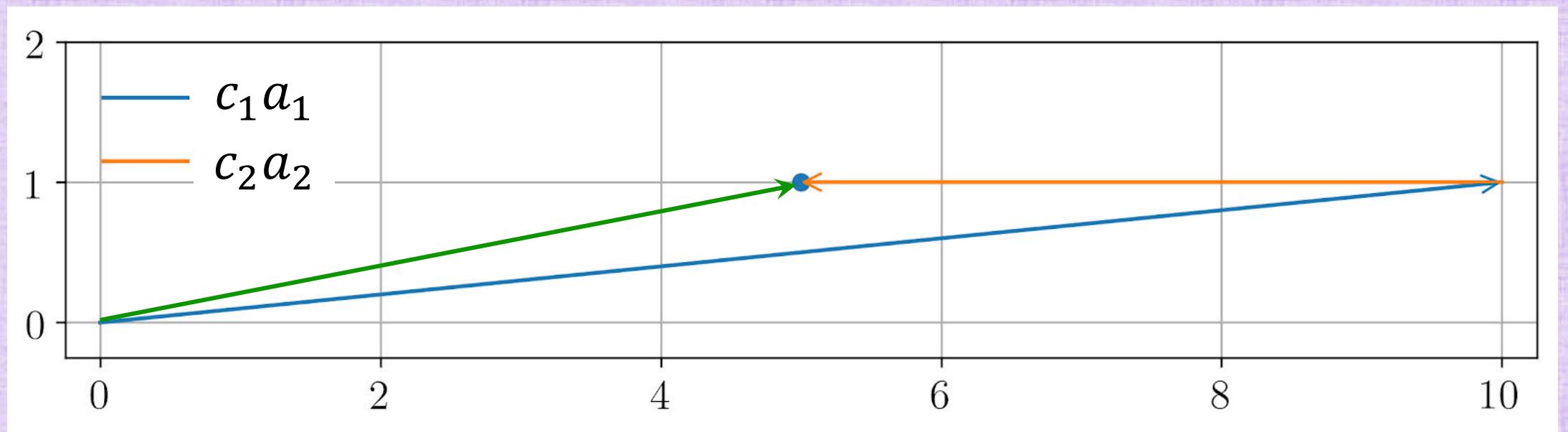
Example

- Determine c_1 and c_2 such that $c_1a_1 + c_2a_2 = b$ or $Ac = b$



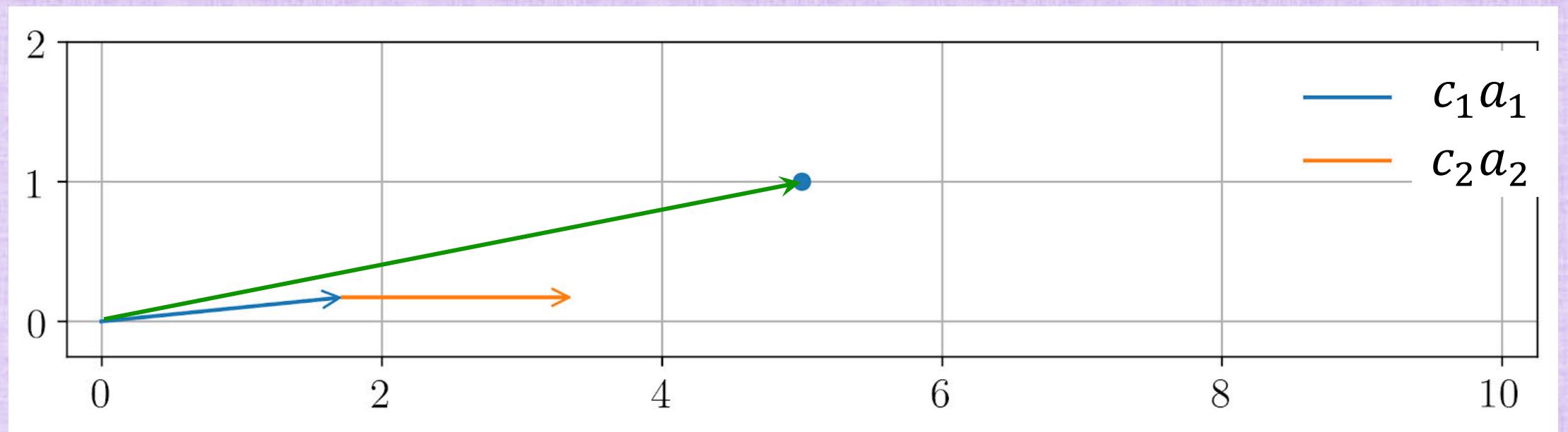
Example (overshooting)

- Since a_1 and a_2 are not parallel, there is a unique solution
- However, this solution overshoots b by quite a bit, and then backtracks



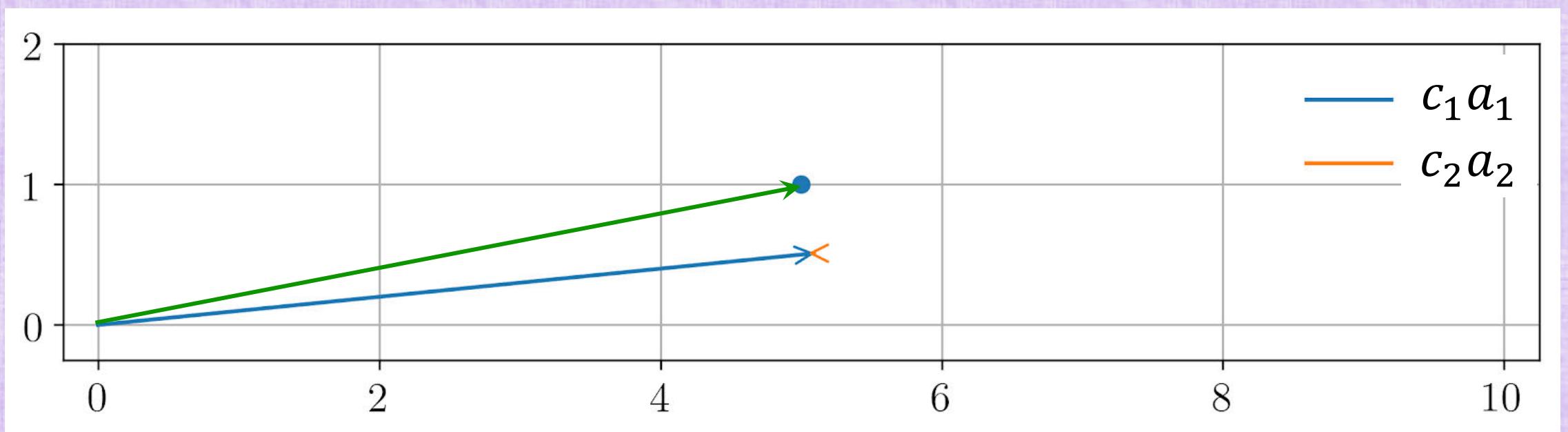
Example (regularization/damping)

- Adding regularization of $Ic = 0$ damps both components of the solution



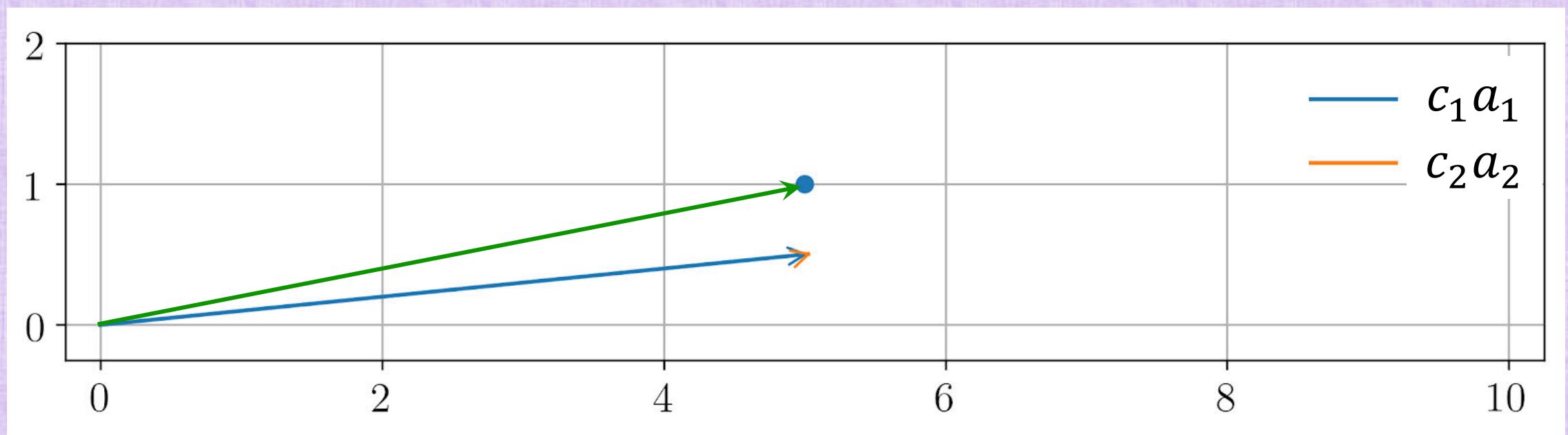
Example (smarter regularization)

- Adding regularization of $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} c = 0$ only damps c_2 and allows $c_1 a_1$ to estimate b unhindered



Example (coordinate descent)

- Coordinate Descent looks at one vector at a time
- After making good progress with a_1 , there is little advantage to using a_2



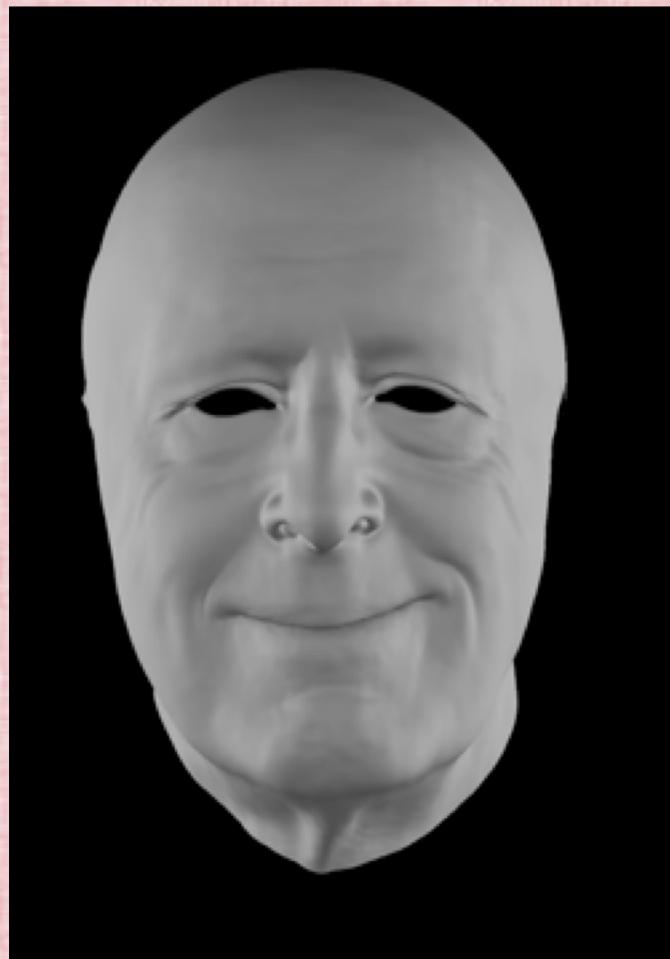
Geometric Approach (Advantages)

- Thinking geometrically avoids issues with the rank of A
- Other concerns may be more important:
 - Use as few columns as possible - Setting many c_k to zero gives a sparser solution (which is easier to glean semantic information from)
 - Correlation - Columns more parallel to b are probably more relevant than those more perpendicular
 - Gains - Columns that have a large dot product with b 's direction make more progress towards b with smaller c_k values (more minimal solution norm)

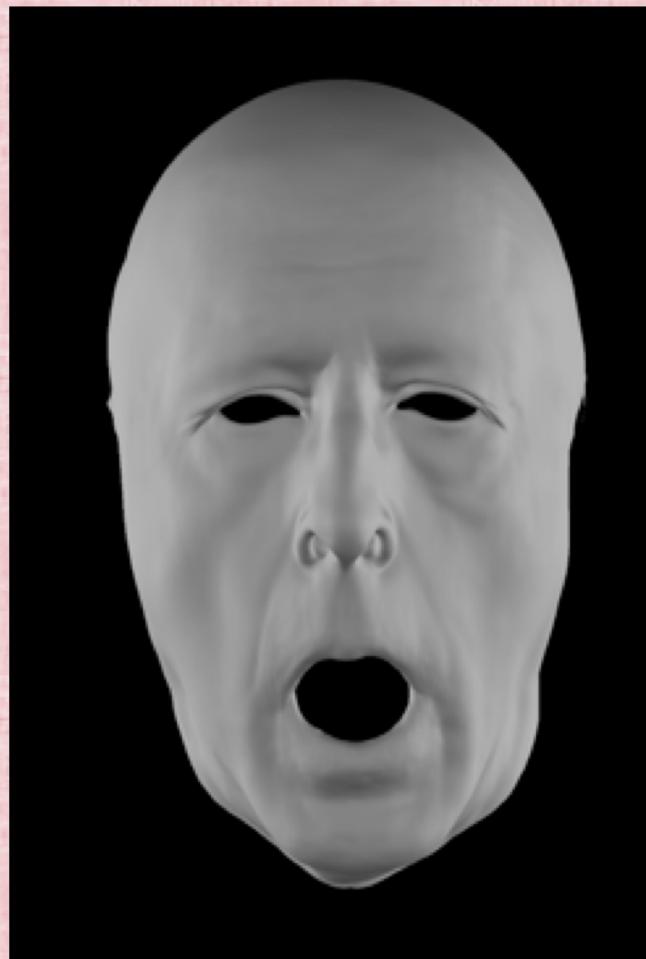
Gains vs. Correlation

- Consider $a_k \cdot b = \|a_k\| \|b\| \cos \theta$ where θ measures how parallel a_k and b are
- Correlation preference uses the columns a_k with smaller θ , i.e. columns that point more closely in the same direction as b
- When the c_k represent actions, the goal of minimizing action/gains leads to a preference for smaller c_k
 - similar in spirit to $Ic = 0$ or minimum norm solutions
- Thus, columns that make more progress in the direction of b are preferable
- Progress in the direction of b is measured via $a_k \cdot \frac{b}{\|b\|}$ or $\|a_k\| \cos \theta$

Facial Animation



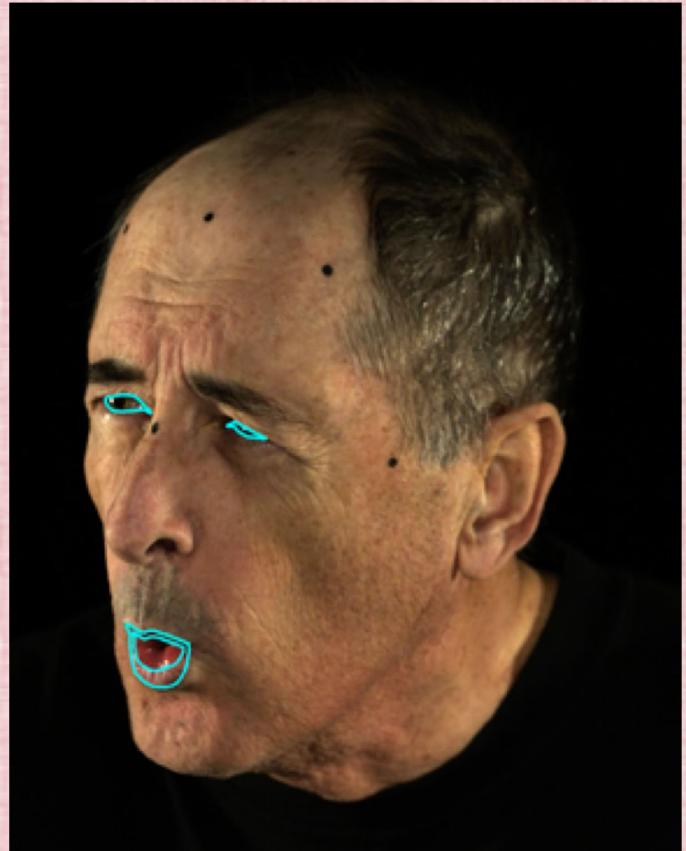
$\varphi(\theta_1)$



$\varphi(\theta_2)$

- Create a procedural skinning model of a face, where (input) animation parameters θ lead to a 3D position (output) for every vertex of the face mesh $\varphi(\theta)$
- E.g. in blend shape systems, each component of θ corresponds to a different expression, and setting multiple components to be nonzero mixes expressions

Facial Tracking



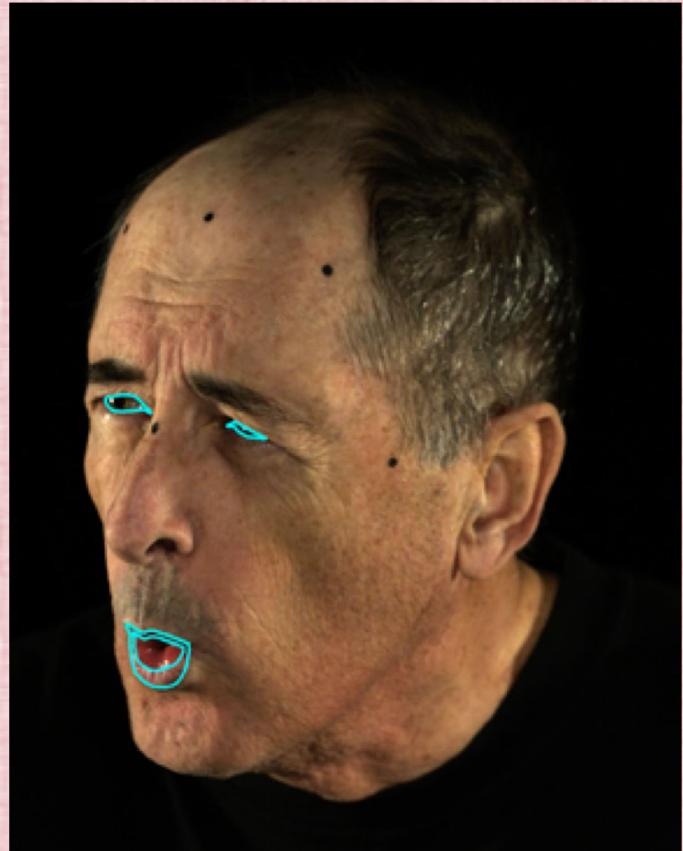
2D RGB Image



3D model

- On the 3D model, embed curves around the eyes/mouth that move with the surface in 3D as it deforms (red curves)
- Draw similar curves on a 2D RGB image of the actual face (blue curves)
- Goal: the projected (onto the image plane) red curves should overlap with the blue curves
 - Once they do, it gives an estimate of the 3D pose in the 2D RGB image

Facial Tracking



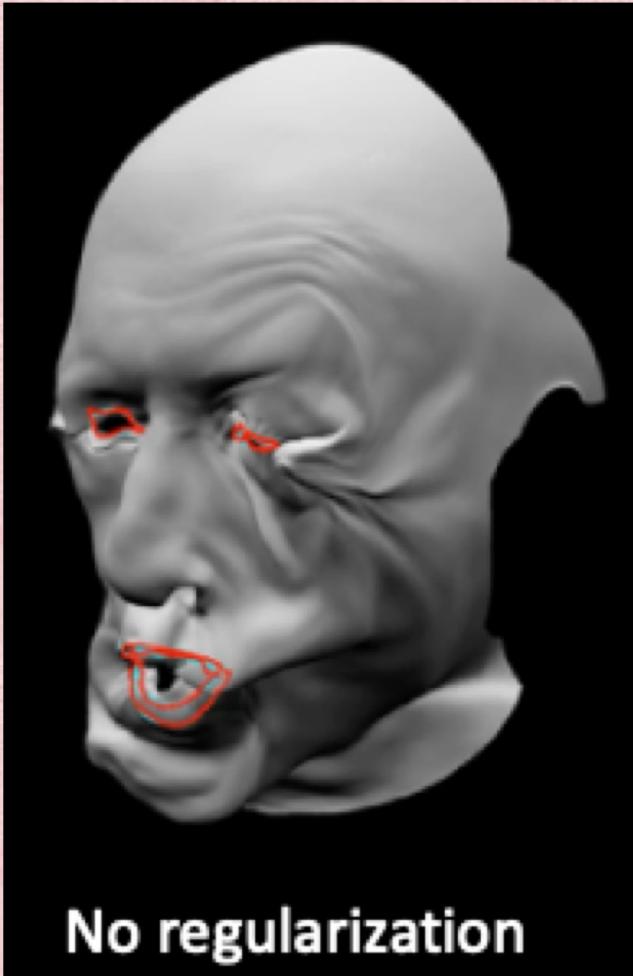
2D RGB Image



3D model

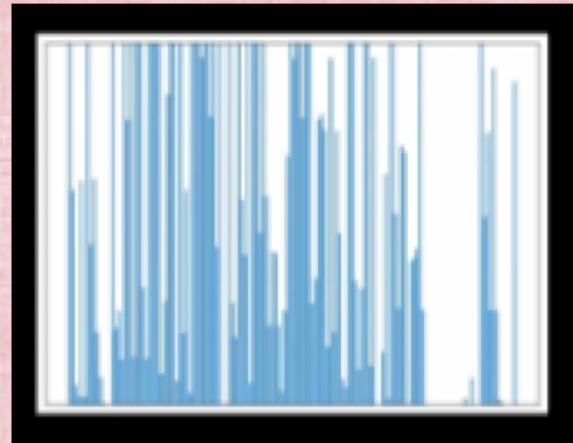
- The blue curves are data C^*
- The red curves C are a function of the 3D geometry φ , which in turn is a function of the animation parameters, i.e. $C(\varphi(\theta))$
- Determine the θ that minimizes the difference $\|C^* - C(\varphi(\theta))\|$ between them

Solving for the Animation Parameters



No regularization

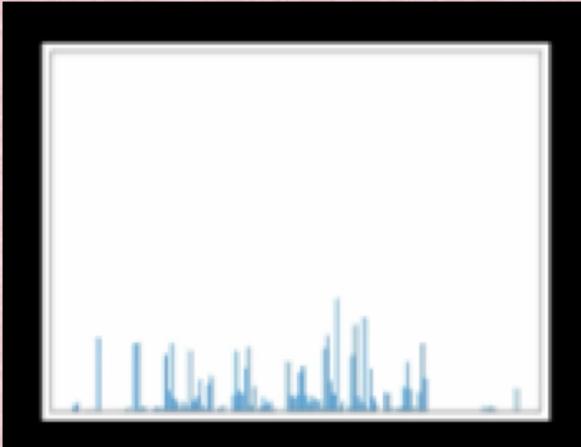
- This is generally a nonlinear problem solved via optimization
- At every step of optimization, the problem is linearized
- The linear problem $Ac = b$ gives a search direction that is used to make progress towards the solution



- The methods performs poorly without regularization
- The resulting θ values are wild and arbitrary (as shown in the figure)

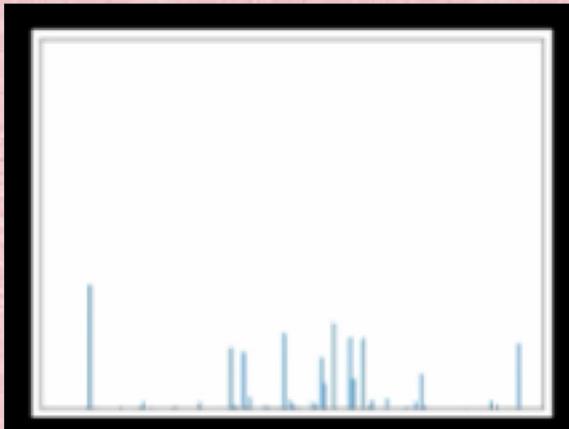
L2 Regularization



- Adding $I\theta = 0$ to the linearized problem at every iteration has the expected result
 - The regularized problem is much more solvable, and the results are less noisy
 - However, θ is overly damped (as seen in the figure)
- 
- Also, a large number of animation parameters θ are nonzero, even though this is a relatively simple expression
 - This hinders interpretability of θ

“Soft L1” Regularization

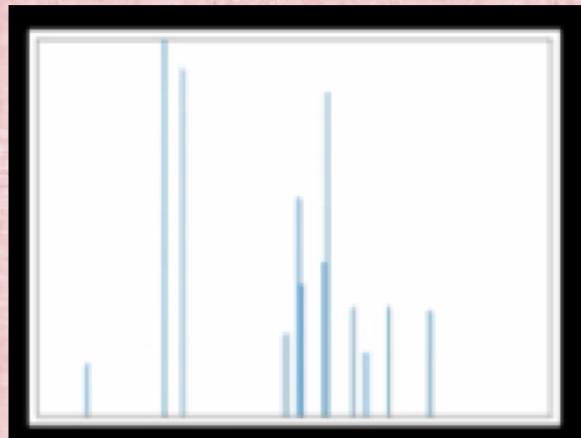
- There are many options for regularization
- In particular, “soft L1” typically produces a sparser set of solution parameters as compared to L2 (see figure)
- A sparser solution allows one to better ascertain semantic meaning from the nonzero θ values
- Although, θ is still overly damped



Soft L1 regularization

Column Space Search

- The column space search gives a sparse set of solution parameters
- A sparser solution allows one to better ascertain semantic meaning from the nonzero θ values
- Moreover, θ is not overly damped



Column Space Search

Unit 13

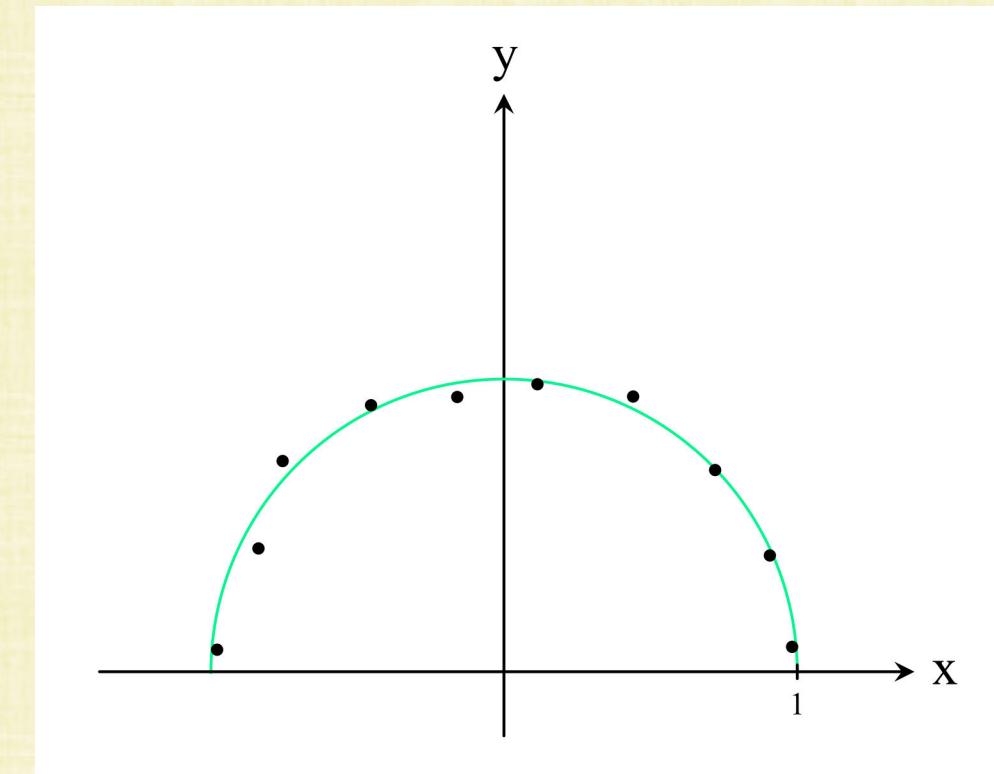
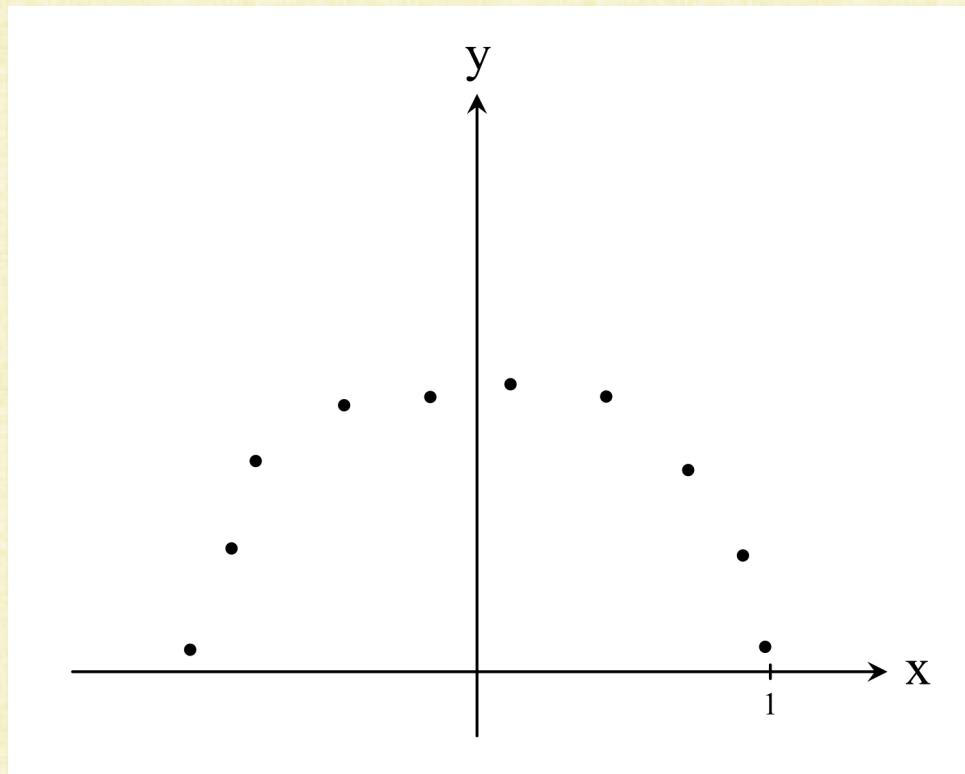
Optimization

Part II Roadmap

- Part I – Linear Algebra (units 1-12) $Ac = b$
 - Part II – Optimization (units 13-20)
 - (units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima
 - (units 17-18) Computing/Avoiding Derivatives
 - (unit 19) Hack 1.0: “I give up” $H = I$ and J is mostly 0 (descent methods)
 - (unit 20) Hack 2.0: “It’s an ODE!?” (adaptive learning rate and momentum)
-
- ```
graph TD; PartI["Part I – Linear Algebra (units 1-12) $Ac = b$ "]; PartII["Part II – Optimization (units 13-20)"]; subgraph Opt ["(units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima"]; direction TB; subgraph Theory ["Theory"]; direction LR; subgraph Methods ["Methods"]; direction LR; end; end; PartI -- "linearize" --> Eq[" $Ac = b$ "]; PartI -- "line search" --> LineSearch["line search"]; Eq --> Opt; Theory --> Opt; Methods --> Opt;
```

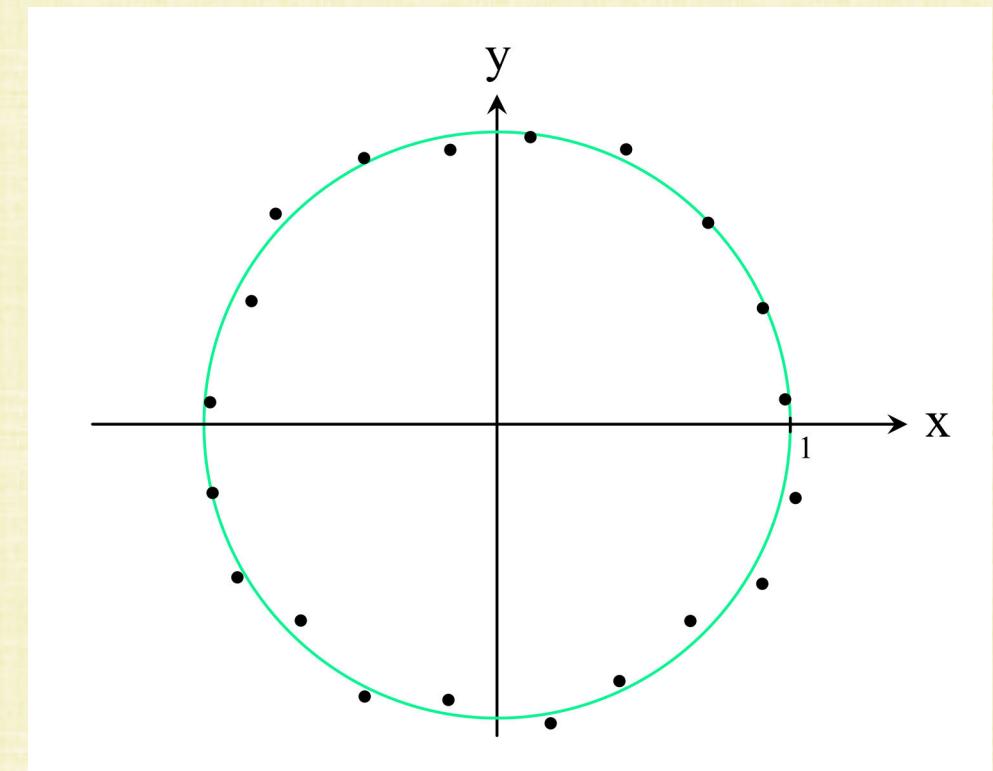
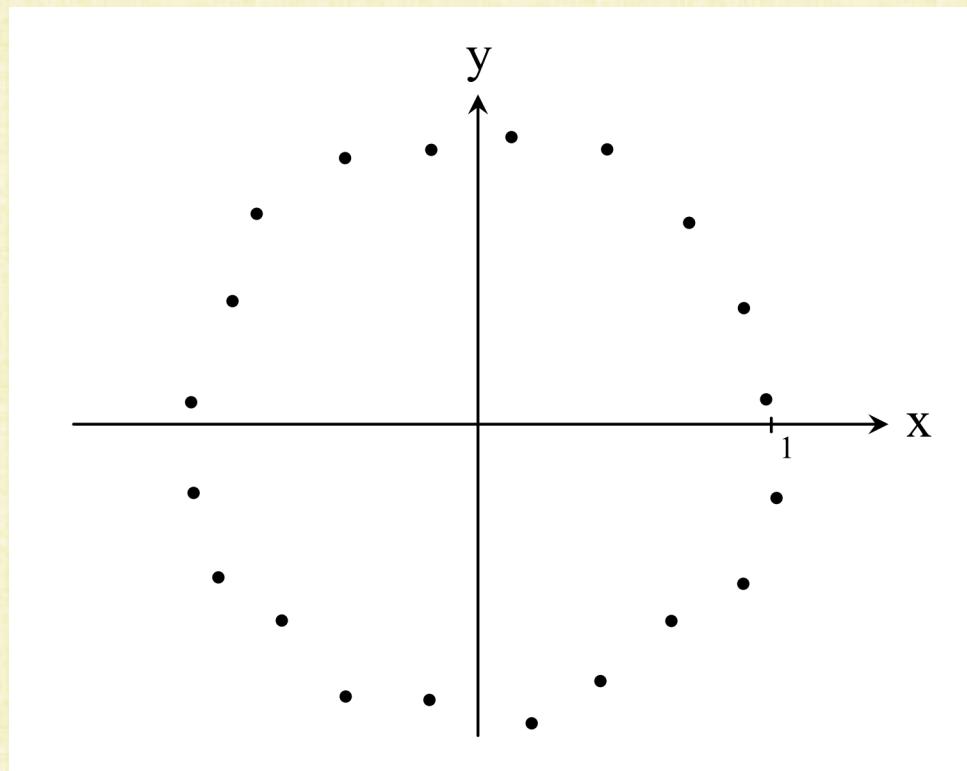
# Function Approximation

- Consider data  $(x_i, y_i)$  as shown below
- Here  $y = \sqrt{1 - x^2}$  looks like a good approximation



# Function Approximation

- Consider data  $(x_i, y_i)$  as shown below
- Here  $x^2 + y^2 = 1$  looks like a good approximation, but it's not a function



# Function Approximation

- More generally, a function need not be explicit in  $y$ , just a general relationship between  $x$  and  $y$ , i.e.  $f(x, y) = 0$
- It is difficult to consider all possible functions at the same time, so one typically chooses a parametric family of possible functions  $f$  (a model for  $f$ )
  - E.g.,  $f$  could be all possible circles  $(x - c_1)^2 + (y - c_2)^3 - c_3 = 0$  where the center  $(c_1, c_2)$  and radius  $c_3$  are chosen to best fit the data
- $f(x, y, c) = 0$  could be a family of polynomials, or circles, or a network architecture, etc.
- Determine parameters  $c$  that make  $f(x, y, c) = 0$  best fit the training data, i.e. that make  $\|f(x_i, y_i, c)\|$  close to zero for all  $i$ 
  - Don't forget to be careful about overfitting/underfitting

# Choosing a Norm

- $f(x, y, c)$  may have scalar or vector output; in the latter case, a norm needs to be chosen for  $\|f(x_i, y_i, c)\|$ , e.g.  $L^1$ ,  $L^2$ ,  $L^\infty$ , “soft”  $L^1$ , etc.
  - E.g.,  $\|f(x_i, y_i, c)\|_2 = \sqrt{f(x_i, y_i, c)^T f(x_i, y_i, c)}$
- There is an  $f(x_i, y_i, c)$  for each ordered pair  $(x_i, y_i)$ , so a norm needs to be chosen to combine these together as well
  - E.g.,  $\sqrt{\sum_i \|f(x_i, y_i, c)\|_2^2} = \sqrt{\sum_i f(x_i, y_i, c)^T f(x_i, y_i, c)}$
- Find  $c$  that minimizes  $\sqrt{\sum_i f(x_i, y_i, c)^T f(x_i, y_i, c)}$ , or equivalently that minimizes  $\sum_i f(x_i, y_i, c)^T f(x_i, y_i, c)$
- Since all the  $(x_i, y_i)$  are known, the cost function is only a function of  $c$ 
  - Minimize  $\hat{f}(c) = \sum_i f(x_i, y_i, c)^T f(x_i, y_i, c)$ , which is Nonlinear Least Squares

# Optimization

- Minimize cost function  $\hat{f}(c)$ , possibly subject to some constraints
- The constraints are equations or inequalities (e.g.  $c_k > 0$  for all  $k$ )
- Constraints can often be folded into the cost function, if one is willing to accept the consequences (more on this later)
- When constraints are present, it is called constrained optimization; otherwise, it is called unconstrained optimization
- Since maximizing  $\hat{f}(c)$  is equivalent to minimizing  $-\hat{f}(c)$ , optimization is typically (always) approached as a minimization problem
- Optimization algorithms often get stuck in and/or only guarantee the ability to find local minima (presumably one might prefer global minima)
  - Sometimes finding lots of local minima, and then choosing the smallest of those, is a good strategy

# Conditioning

- Recall: Minimizing the residual  $r = b - Ac$  led to normal equations  $A^T Ac = A^T b$  that square the condition number
- This is an issue for optimization as well:
  - Optimization considers critical points where  $\frac{\partial \hat{f}}{\partial c_k}(c) = 0$  simultaneously for all  $k$
  - Having all partial derivatives approach zero near critical points makes the function locally flat, and thus algorithms struggle to find robust downhill search directions
- The condition number for minimizing  $\hat{f}(c)$  is typically the square of that for solving  $\hat{f}(c) = 0$ 
  - Can only expect half as many significant digits of accuracy
  - If an error tolerance of  $\epsilon$  would be used for solving  $\hat{f}(c) = 0$ , then  $\sqrt{\epsilon}$  is more appropriate for minimizing  $\hat{f}(c)$

# Nonlinear Systems of Equations

- The critical points are the points where  $\frac{\partial \hat{f}}{\partial c_k}(c) = 0$  simultaneously for all  $k$
- Stacking all the (potentially) nonlinear functions  $\frac{\partial \hat{f}}{\partial c_k}(c)$  into a single vector valued function, the critical points are solutions to:  $F(c) = \begin{pmatrix} \frac{\partial \hat{f}}{\partial c_1}(c) \\ \frac{\partial \hat{f}}{\partial c_2}(c) \\ \vdots \\ \frac{\partial \hat{f}}{\partial c_n}(c) \end{pmatrix} = 0$
- $F(c) = J_{\hat{f}}^T(c) = \nabla \hat{f}(c) = 0$  is a nonlinear system of equations
  - It may have no solution, any finite number of solutions, or infinite solutions

# (Equality) Constrained Optimization

- Constraints can be equalities, e.g.  $\hat{g}(c) = 0$ , or inequalities (see unit 17)
- Given a diagonal matrix  $D$  of (positive) weights indicating the relative importance of various constraints, one can add a penalty term of the form  $\hat{g}^T(c)D\hat{g}(c) \geq 0$  to the cost function and proceed with unconstrained optimization
  - I.e., minimize  $\hat{f}(c) + \hat{g}^T(c)D\hat{g}(c)$  via unconstrained optimization
- Various other options also exist
- An alternative approach uses Lagrange multipliers  $\eta$  as new variables, and minimizes  $\hat{f}(c) + \eta^T \hat{g}(c)$

# Lagrange Multipliers

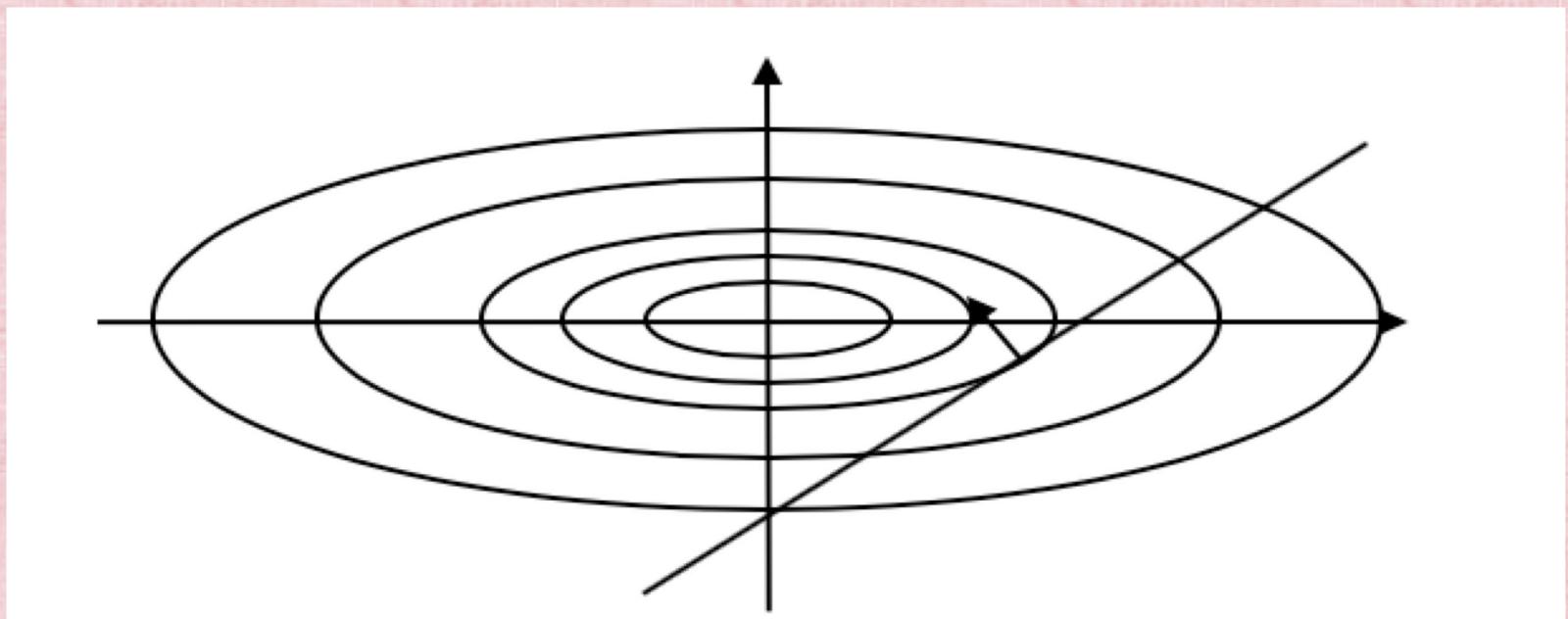
- Minimize  $\hat{f}(c) + \eta^T \hat{g}(c)$
- Critical Points:  $\nabla(\hat{f}(c) + \eta^T \hat{g}(c)) = \begin{pmatrix} J_{\hat{f}}^T(c) + (\eta^T J_{\hat{g}}(c))^T \\ \hat{g}(c) \end{pmatrix} = 0$ 
  - So the constraints  $\hat{g}(c) = 0$  are automatically satisfied
- Critical points satisfy  $J_{\hat{f}}^T(c) = -J_{\hat{g}}^T(c)\eta$  instead of the usual  $J_{\hat{f}}^T(c) = 0$
- In the simple case when  $\hat{g}(c)$  is linear in  $c$ , the Hessian is  $\begin{pmatrix} H_{\hat{f}}(c) & J_{\hat{g}}^T \\ J_{\hat{g}} & 0 \end{pmatrix}$  which is symmetric but not positive definite
  - However, positive definiteness is only required on the tangent space to the constraint surface (i.e., on the null space of  $J_{\hat{g}}$ )

# Lagrange Multipliers (Example)

- Minimize  $\hat{f}(c) = .5c_1^2 + 2.5c_2^2$  subject to  $\hat{g}(c) = c_1 - c_2 - 1 = 0$
- So, minimize  $.5c_1^2 + 2.5c_2^2 + \eta(c_1 - c_2 - 1)$
- Critical Points:  $\begin{pmatrix} c_1 \\ 5c_2 \\ c_1 - c_2 - 1 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}\eta = \begin{pmatrix} c_1 + \eta \\ 5c_2 - \eta \\ c_1 - c_2 - 1 \end{pmatrix} = 0$
- $\begin{pmatrix} 1 & 0 & 1 \\ 0 & 5 & -1 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \eta \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$  or  $\begin{pmatrix} c_1 \\ c_2 \\ \eta \end{pmatrix} = \begin{pmatrix} 5/6 \\ -1/6 \\ -5/6 \end{pmatrix}$
- The Hessian is  $\begin{pmatrix} (1 & 0) & (1) \\ (0 & 5) & (-1) \\ (1 & -1) & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 5 & -1 \\ 1 & -1 & 0 \end{pmatrix}$

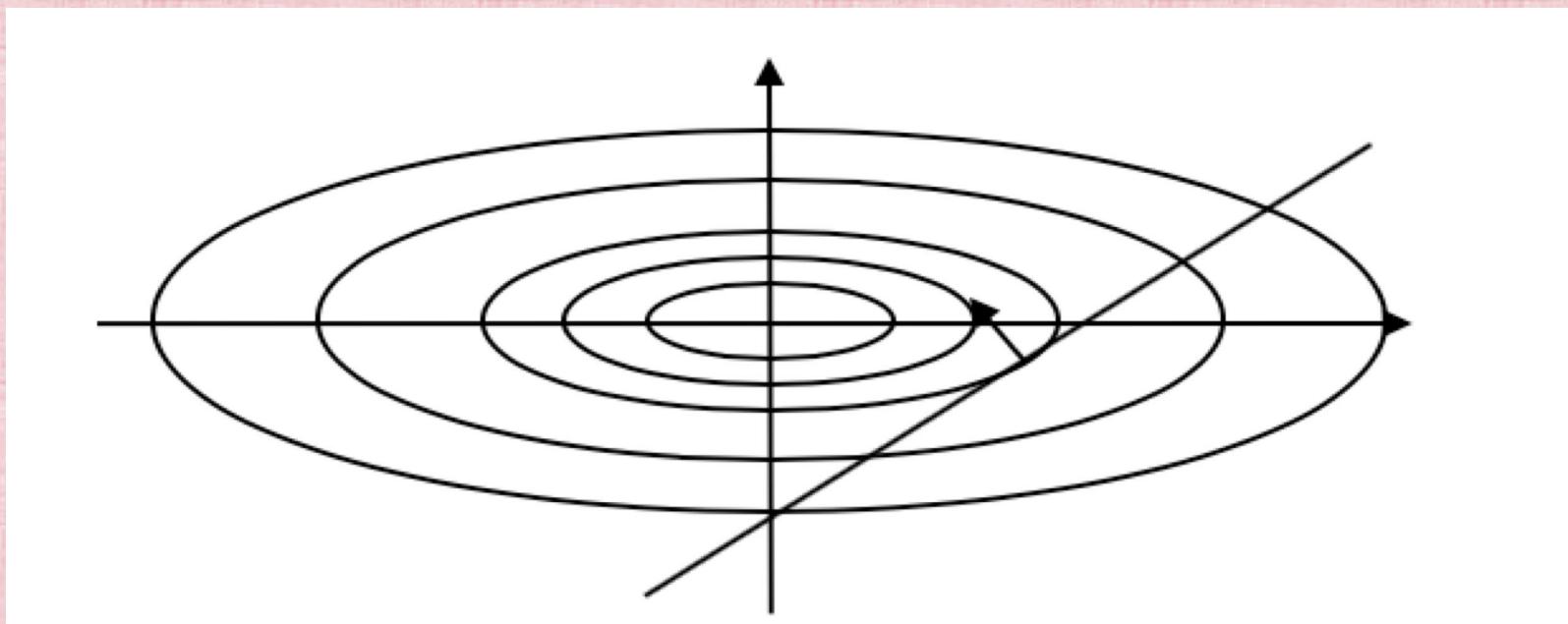
# Lagrange Multipliers (Example)

- Isocontours of  $\hat{f}(c)$  are ellipses, and the constraint is the line  $c_2 = c_1 - 1$
- At critical point  $\left(\frac{5}{6}, -\frac{1}{6}\right)$ , the steepest descent direction  $-\nabla \hat{f} = \left(-\frac{5}{6}, \frac{5}{6}\right)$  is perpendicular to the constraint surface that has direction  $(1,1)$



# Lagrange Multipliers (Example)

- Plugging  $c_2 = c_1 - 1$  into  $\hat{f}(c)$  gives  $.5c_1^2 + 2.5(c_1 - 1)^2 = 3c_1^2 - 5c_1 + 2.5$  which is a parabola with minimum at  $c_1 = \frac{5}{6}$  (as expected)



# Unit 14

# Nonlinear Systems

# Part II Roadmap

- Part I – Linear Algebra (units 1-12)  $Ac = b$ 
    - Part II – Optimization (units 13-20)
      - (units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima
      - (units 17-18) Computing/Avoiding Derivatives
      - (unit 19) Hack 1.0: “I give up”  $H = I$  and  $J$  is mostly 0 (descent methods)
      - (unit 20) Hack 2.0: “It’s an ODE!?” (adaptive learning rate and momentum)
- 
- ```
graph TD; A["Part I – Linear Algebra (units 1-12)  $Ac = b$ "] -- "linearize" --> B["Nonlinear Equations"]; A -- "line search" --> C["1D roots/minima"]; D["Part II – Optimization (units 13-20)"] --- E["(units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima"]; D --- F["(units 17-18) Computing/Avoiding Derivatives"]; D --- G["(unit 19) Hack 1.0: ‘I give up’  $H = I$  and  $J$  is mostly 0 (descent methods)"]; D --- H["(unit 20) Hack 2.0: ‘It’s an ODE!?’ (adaptive learning rate and momentum)"]
```

Recall: Jacobian

- Given $F(c) = \begin{pmatrix} F_1(c) \\ F_2(c) \\ \vdots \\ F_m(c) \end{pmatrix}$ the Jacobian of $F(c)$ has entries $J_{ij} = \frac{\partial F_i}{\partial c_j}(c)$
- Thus, the Jacobian $J(c) = F'(c) = \begin{pmatrix} \frac{\partial F_1}{\partial c_1}(c) & \frac{\partial F_1}{\partial c_2}(c) & \cdots & \frac{\partial F_1}{\partial c_n}(c) \\ \frac{\partial F_2}{\partial c_1}(c) & \frac{\partial F_2}{\partial c_2}(c) & \cdots & \frac{\partial F_2}{\partial c_n}(c) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial c_1}(c) & \frac{\partial F_m}{\partial c_2}(c) & \cdots & \frac{\partial F_m}{\partial c_n}(c) \end{pmatrix}$

Linearization

- Solving the nonlinear system $F(c) = 0$ is difficult
- It can be linearized by considering the first term in the multidimensional version of the Taylor expansion: $F(c) \approx F(c^*) + F'(c)(c - c^*) = F(c^*) + F'(c)\Delta c$
 - This is more valid when Δc is small (i.e. for c close enough to c^*)
 - This can be alternatively written as $F(c) - F(c^*) \approx F'(c)\Delta c$
- The chain rule $\frac{dF(c)}{dt} = F'(c) \frac{dc}{dt}$ can be written in differential form for vanishingly small differentials $dF(c) = F'(c)dc$
 - This is often referred to as the total derivative
 - Using finite size differentials leads to the approximation $\Delta F(c) \approx F'(c)\Delta c$
 - In 1D, $df = f'(c)dc$ and $\Delta f \approx f'(c)\Delta c$ are the usual $\frac{df}{dc} = f'(c)$ and $\frac{\Delta f}{\Delta c} \approx f'(c)$

Newton's Method

- Iteratively, starting with c^0 , recursively find: c^1, c^2, c^3, \dots
- Newton's Method uses $\Delta F(c) \approx F'(c)\Delta c$ to write $F(c^{q+1}) - F(c^q) = F'(c^q)\Delta c^q$ where $\Delta c^q = c^{q+1} - c^q$
 - Aiming for $F(c) = 0$ motivates setting $F(c^{q+1}) = 0$
 - Alternatively, one could set $F(c^{q+1}) = \beta F(c^q)$ where $0 \leq \beta < 1$ gives a rate at which one aims to shrink $F(c^q)$ towards zero
 - This gives $F'(c^q)\Delta c^q = (\beta - 1)F(c^q)$ with β often set to 0
- The linear system $F'(c^q)\Delta c^q = (\beta - 1)F(c^q)$ is solved for Δc^q , which is used to update $c^{q+1} = c^q + \Delta c^q$

Newton's Method

- Requires repeated solving of a linear system, which is one reason why robustness and efficiency for linear system solvers is so important
 - Need to consider size, rank, conditioning, symmetry, etc. of $F'(c^q)$
- $F'(c^q)$ may be difficult to compute, since it requires every first derivative
 - In particular, Newton's Method contains linearization errors, so useful approximations of $F'(c^q)$ seem valid/worthwhile (e.g. symmetrization, etc.)
 - This is discussed more in units 17/18
- Generally, there are no guarantees on convergence
 - May converge to any one of many roots when multiple roots exist or not converge at all

Linear System Solvers (Review)

- Theory, all matrices: **SVD** (unit 3/9/11)
- Square, full rank, dense:
 - LU factorization with pivoting (unit 2)
 - SPD: **Cholesky** factorization (unit 4), **Symmetric approximation** (unit 4)
- Square, full rank, sparse (iterative solvers) (unit 5):
 - SPD (sometimes SPSD): **Conjugate Gradients**
 - Nonsymmetric/Indefinite: GMRES, MINRES, BiCGSTAB (not steepest descent)
- Tall, full rank (least squares to minimize residual) (unit 8):
 - normal equations (unit 9/10), **QR**, Gram-Schmidt, **Householder** (unit 10)
- Any size/rank (minimum norm solution) (unit 11):
 - Pseudo-Inverse, **PCA approximation**, **Power Method** (unit 11)
 - **Levenberg-Marquardt** (iteration too), **Column Space Geometric Approach** (unit 12)

Line Search

- Given the linearization error in $F'(c^q)\Delta c^q = -F(c^q)$, the resulting Δc^q often leads to a poor estimate for c^{q+1} via $c^{q+1} = c^q + \Delta c^q$
- Thus, Δc^q is often merely treated as a search direction, i.e. $c^{q+1} = c^q + \alpha^q \Delta c^q$
- The parameterized line $c^{q+1}(\alpha) = c^q + \alpha \Delta c^q$ is used as a 1D (input) domain
- Find α such that $F(c^{q+1}(\alpha)) = 0$ simultaneously for all equations
- Safe Set methods restrict α in various ways, e.g. $0 \leq \alpha \leq 1$

Line Search

- Since F is vector valued, consider $g(\alpha) = F(c^{q+1}(\alpha))^T F(c^{q+1}(\alpha)) = 0$
- Note $g(\alpha) = F(c^{q+1}(\alpha))^T F(c^{q+1}(\alpha)) \geq 0$, so solutions to $F(c^{q+1}(\alpha)) = 0$ are minima of $g(\alpha)$ (and more difficult to find)
- $g(\alpha)$ might be strictly positive, but minimizing $g(\alpha)$ can help to make progress towards a solution
- Option 1: find simultaneous roots of the vector valued $F(c^{q+1}(\alpha)) = 0$
- Option 2: find roots or minimize $g(\alpha) = \frac{1}{2} F^T(c^{q+1}(\alpha)) F(c^{q+1}(\alpha))$

Optimization Problems

- Minimize the scalar cost function $\hat{f}(c)$ by finding the critical points where $\nabla \hat{f}(c) = J_{\hat{f}}^T(c) = F(c) = 0$
- $F'(c^q)\Delta c^q = -F(c^q)$ gives the search direction, where $F'(c) = J_F(c) = H_{\hat{f}}^T(c)$
- That is, solve $H_{\hat{f}}^T(c^q)\Delta c^q = -J_{\hat{f}}^T(c^q)$ to find the search direction Δc^q
- Option 1: find simultaneous roots of the vector valued $J_{\hat{f}}^T(c^{q+1}(\alpha)) = 0$ which are the critical points of $\hat{f}(c)$
- Option 2: find roots or minimize $g(\alpha) = \frac{1}{2}J_{\hat{f}}(c^{q+1}(\alpha))J_{\hat{f}}^T(c^{q+1}(\alpha))$ to find or make progress toward critical points of $\hat{f}(c)$
- Option 3: minimize $\hat{f}(c^{q+1}(\alpha))$ directly

Unit 15

1D Root Finding

Part II Roadmap

- Part I – Linear Algebra (units 1-12) $Ac = b$
 - Part II – Optimization (units 13-20)
 - (units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima
 - (units 17-18) Computing/Avoiding Derivatives
 - (unit 19) Hack 1.0: “I give up” $H = I$ and J is mostly 0 (descent methods)
 - (unit 20) Hack 2.0: “It’s an ODE!?” (adaptive learning rate and momentum)
-
- The diagram illustrates the connection between Part I and Part II. A red arrow labeled "linearize" points from the text "Part I – Linear Algebra" to the equation $Ac = b$. Another red arrow labeled "line search" points from the text "Part I – Linear Algebra" to the term "1D roots/minima". On the right side of Part II, a blue bracket groups the four items under the labels "Theory" (above the bracket) and "Methods" (below the bracket).

Fixed Point Iteration

- Find roots of $g(t)$ where $g(t) = 0$
- Let $\hat{g}(t) = g(t) + t$ and iterate $t^{q+1} = \hat{g}(t^q)$ until convergence
- The converged t^* satisfies $t^* = \hat{g}(t^*) = g(t^*) + t^*$, and so $g(t^*) = 0$
- Converges when $|g'(t^*)| < 1$ for a close enough initial guess (when g is sufficiently smooth)
- $e^{q+1} = t^{q+1} - t^* = \hat{g}(t^q) - \hat{g}(t^*) = g'(\hat{t})(t^q - t^*) = g'(\hat{t})e^q$ for some \hat{t} between t^{q+1} and t^* (by the Mean Value Theorem)
- When all $g'(\hat{t})$ have $|g'(\hat{t})| \leq C < 1$, then $|e^q| \leq C^q |e^0|$ proves convergence

Convergence Rate

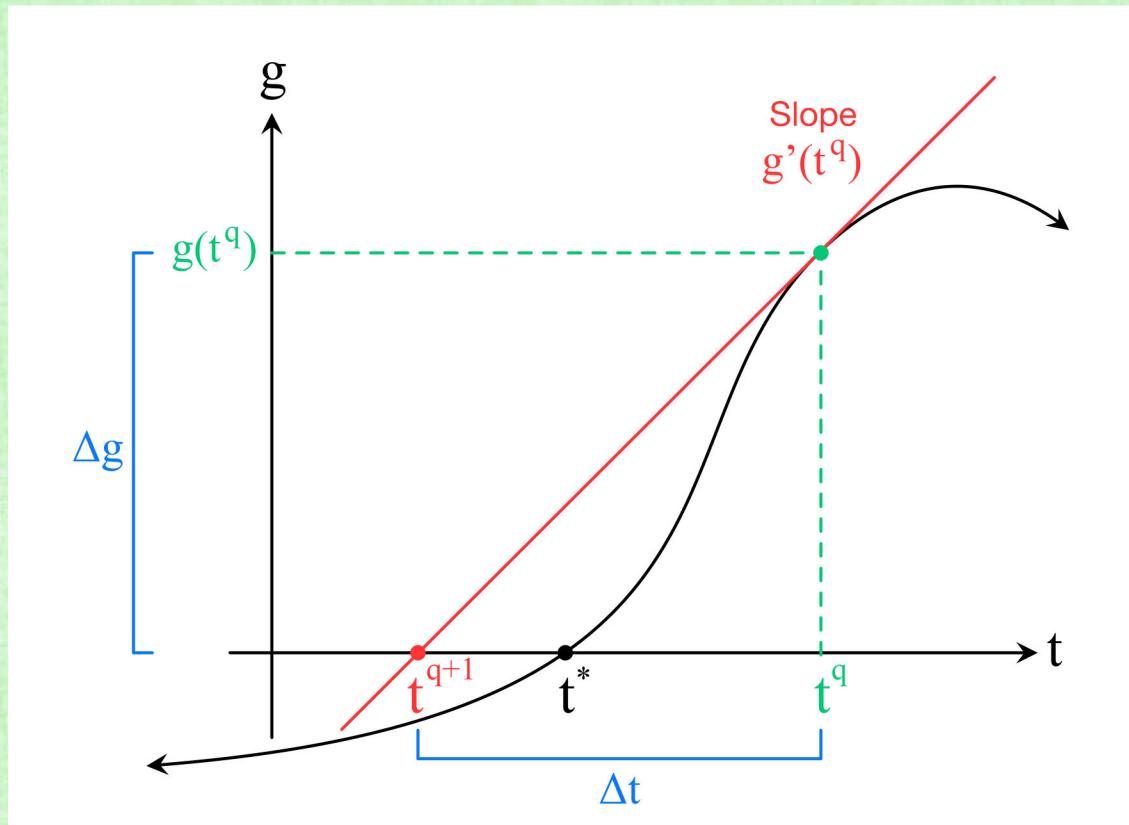
- Consider $\|e^{q+1}\| \leq C\|e^q\|^p$ as $q \rightarrow \infty$ where $C \geq 0$
 - When $p = 1$, $C < 1$ is required, and the convergence rate is linear
 - When $p > 1$, the convergence rate is superlinear
 - When $p = 2$, the convergence rate is quadratic
- Statements only apply asymptotically (once convergence is happening)
- No guarantee of converging to the desired root (when others are present)
- Recall, $g(t) = 0$ may contain approximations, so it's not clear how accurate the root finder needs to be

1D Newton's Method

- Solve $g'(t^q)\Delta t = -g(t^q)$ and update $t^{q+1} = t^q + \Delta t = t^q - \frac{g(t^q)}{g'(t^q)}$
- Stop when $|g(t^q)| < \epsilon$, which implies $|t^{q+1} - t^q| < \frac{\epsilon}{|g'(t^q)|}$
 - Thus, poorly conditioned when $g'(t^*)$ is small
 - Especially problematic for multiple roots where $g'(t^*) = 0$
- Quadratic convergence rate ($p = 2$)
- Requires computing g and g' every iteration, and computing derivatives isn't always straightforward/cheap (see units 17/18)

1D Newton's Method

- $t^{q+1} = t^q - \frac{g(t^q)}{g'(t^q)}$ or alternatively $g'(t^q) = \frac{\Delta g}{\Delta t} = \frac{g(t^q) - 0}{t^q - t^{q+1}}$

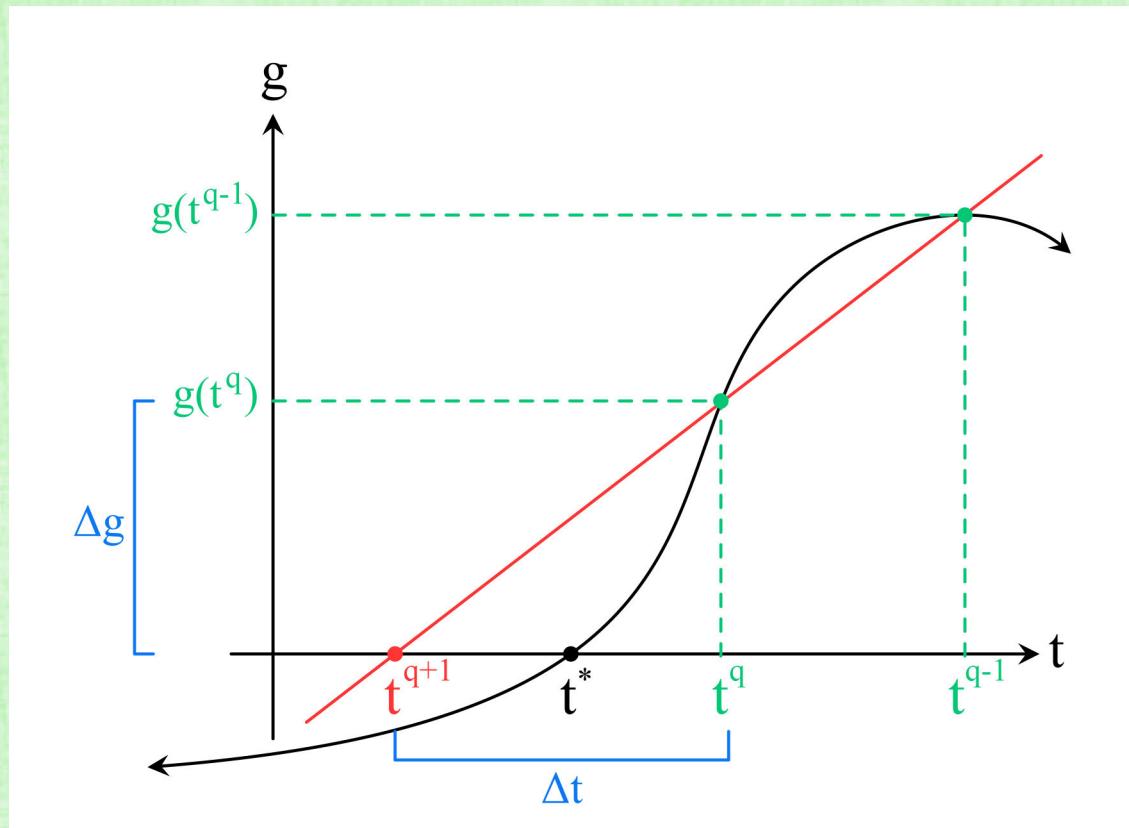


Secant Method

- Replace $g'(t^q)$ in Newton's method with an estimate (a few choices for this)
- Standard technique/method draws a line through previous iterates
- Estimate $g'(t^q) \approx \frac{g(t^q) - g(t^{q-1})}{t^q - t^{q-1}}$
- Then $t^{q+1} = t^q - g(t^q) \frac{t^q - t^{q-1}}{g(t^q) - g(t^{q-1})}$
- Superlinear convergence rate with $p \approx 1.618$
- Often/typically faster than Newton, since only g (not g') is needed while only a few extra iterations are required for the same accuracy

Secant Method

- $t^{q+1} = t^q - g(t^q) \frac{t^q - t^{q-1}}{g(t^q) - g(t^{q-1})}$ based on $g'(t^q) \approx \frac{g(t^q) - g(t^{q-1})}{t^q - t^{q-1}}$

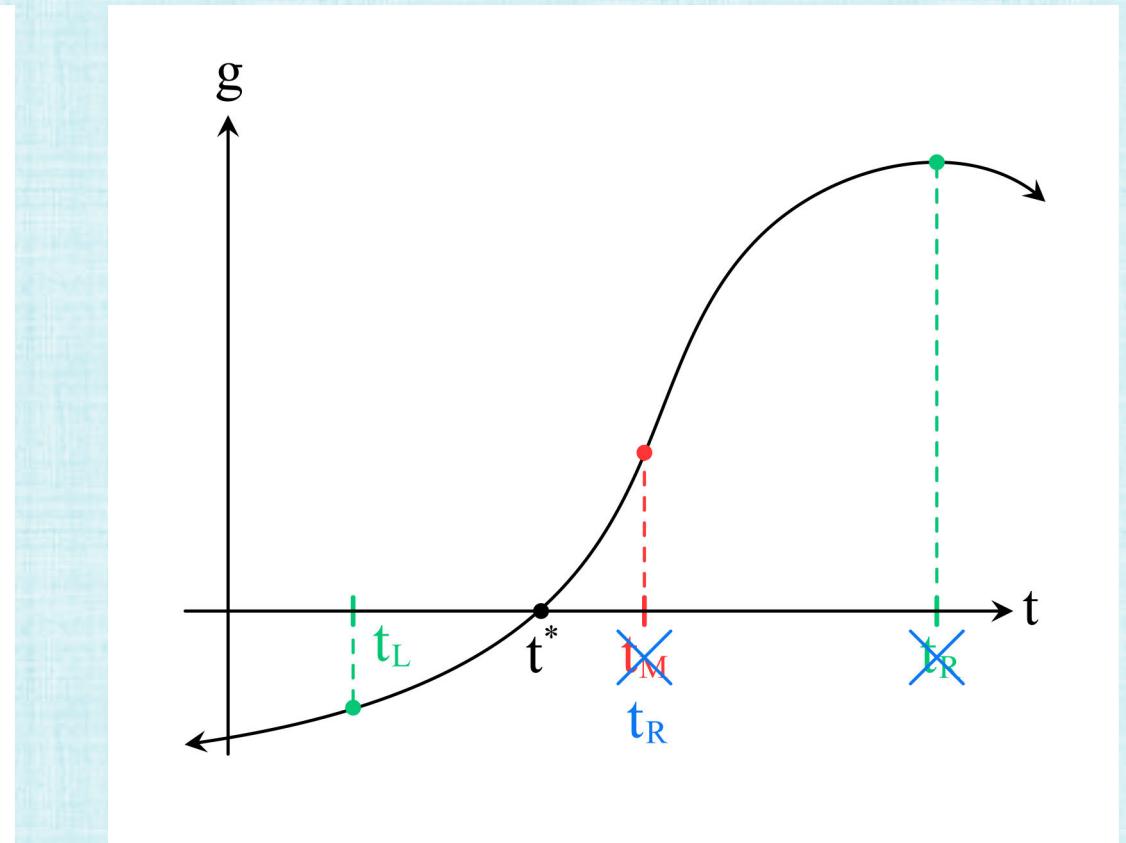
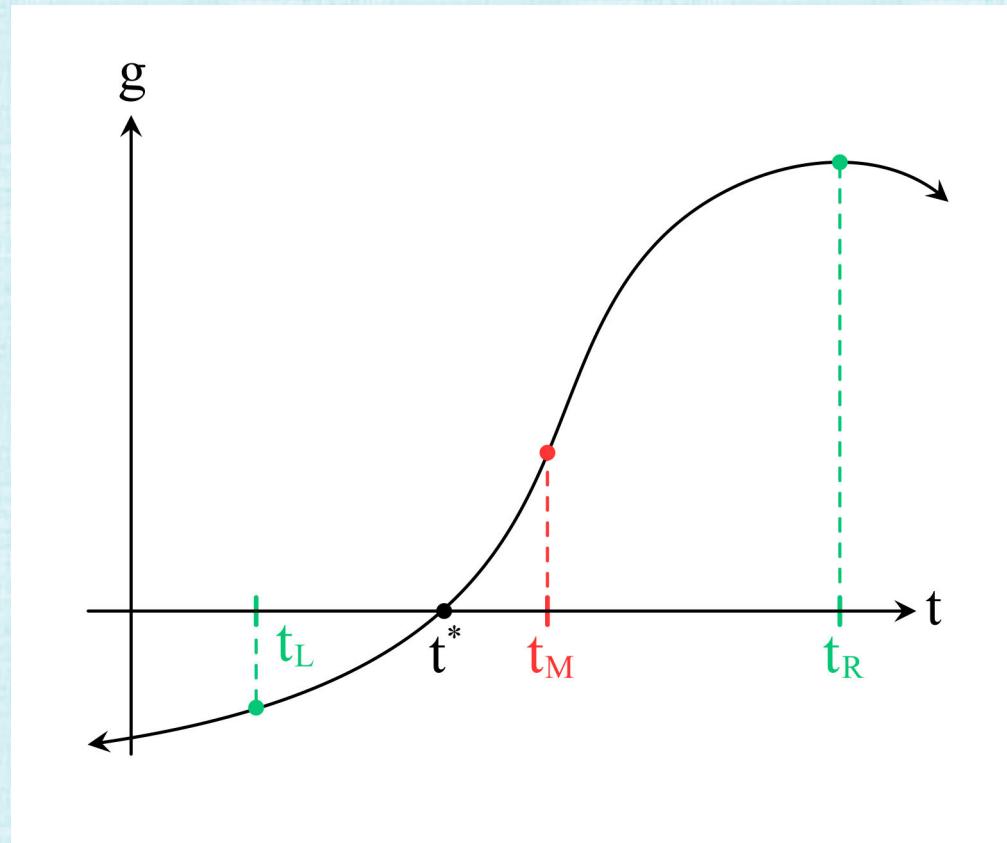


Bisection Method

- If $g(t_L)g(t_R) < 0$ then (when continuous) the sign change indicates a root in the interval $[t_L, t_R]$
- Let $t_M = \frac{t_L + t_R}{2}$, and if $g(t_L)g(t_M) < 0$, set $t_R = t_M$
 - Otherwise, set $t_L = t_M$ knowing that $g(t_M)g(t_R) < 0$ is true
- Iterate until $t_R - t_L < \epsilon$
- Guaranteed to converge to a root in the interval (unlike Newton/Secant)
- The interval shrinks in size by a factor of two each iteration
- So, linear convergence rate ($p = 1$) with $C = \frac{1}{2}$

Bisection Method

- If $g(t_L)g(t_M) < 0$, set $t_R = t_M$; otherwise, set $t_L = t_M$



Mixed Methods

- Given an interval with a root indicated by $g(t_L)g(t_R) < 0$
- Iterate with Newton/Secant as long as the iterates stay inside the interval
 - When iteration attempts to leave the interval, use prior iterates to shrink the interval as much as possible (while still guaranteeing a root)
- Bisection can be used to continue to shrink the interval, whenever Newton/Secant would fail to stay inside the current interval
- Leverages the speed of Newton/Secant, while still guaranteeing convergence via Bisection
- Many/various strategies exist

Function/Derivative Requirements

- All methods require function evaluation g
- Newton requires the derivative g' (as do mixed methods using Newton)

Useful Derivatives

- $\frac{\partial}{\partial t} c^{q+1}(t) = \Delta c^q$, since $c^{q+1}(t) = c^q + t\Delta c^q$
- $\frac{\partial}{\partial t} F(c^{q+1}(t)) = J_F(c^{q+1}(t)) \Delta c^q$ and $\frac{\partial}{\partial t} F^T(c^{q+1}(t)) = (\Delta c^q)^T J_F^T(c^{q+1}(t))$
 - $\frac{\partial}{\partial t} F_i(c^{q+1}(t)) = (J_F)_i(c^{q+1}(t)) \Delta c^q$ where $F_i(c^{q+1}(t))$ are the scalar row components of $F(c^{q+1}(t))$
- Scalar $\hat{f}(c^{q+1}(t))$ has system $J_{\hat{f}}^T(c^{q+1}(t)) = 0$ for critical points
- $\frac{\partial}{\partial t} J_{\hat{f}}^T(c^{q+1}(t)) = H_{\hat{f}}^T(c^{q+1}(t)) \Delta c^q$ and $\frac{\partial}{\partial t} J_{\hat{f}}(c^{q+1}(t)) = (\Delta c^q)^T H_{\hat{f}}(c^{q+1}(t))$
 - $\frac{\partial}{\partial t} (J_{\hat{f}}^T)_i(c^{q+1}(t)) = (H_{\hat{f}}^T)_i(c^{q+1}(t)) \Delta c^q$

Nonlinear Systems Problems

- Solve $J_F(c^q)\Delta c^q = -F(c^q)$ for Δc^q and use $c^{q+1}(t) = c^q + t\Delta c^q$ in $F(c^{q+1}(t)) = 0$
- Option 1: For vector valued $F(c^{q+1}(t))$, find simultaneous (for all i) **roots** for all the $g_i(t) = F_i(c^{q+1}(t)) = 0$
 - Here, $g'_i(t) = (J_F)_i(c^{q+1}(t))\Delta c^q$
- Option 2: Find **roots** of $g(t) = \frac{1}{2}F^T(c^{q+1}(t))F(c^{q+1}(t)) = 0$
 - Here, $g'(t) = \frac{1}{2}F^T(c^{q+1}(t))J_F(c^{q+1}(t))\Delta c^q + \frac{1}{2}(\Delta c^q)^T J_F^T(c^{q+1}(t))F(c^{q+1}(t))$
 - Both terms are scalars, so $g'(t) = F^T(c^{q+1}(t))J_F(c^{q+1}(t))\Delta c^q$

Optimization Problems

- Solve $H_{\hat{f}}^T(c^q)\Delta c^q = -J_{\hat{f}}^T(c^q)$ for Δc^q and use $c^{q+1}(t) = c^q + t\Delta c^q$ in $J_{\hat{f}}^T(c^{q+1}(t)) = 0$
- Option 1: For vector valued $J_{\hat{f}}^T(c^{q+1}(t))$, find simultaneous (for all i) **roots** for all the $g_i(t) = (J_{\hat{f}}^T)_i(c^{q+1}(t)) = 0$ to find the critical points of $\hat{f}(c)$
 - Here, $g'_i(t) = (H_{\hat{f}}^T)_i(c^{q+1}(t))\Delta c^q$
- Option 2: Find **roots** of $g(t) = \frac{1}{2}J_{\hat{f}}^T(c^{q+1}(t))J_{\hat{f}}(c^{q+1}(t)) = 0$ to find or make progress toward critical points of $\hat{f}(c)$
 - Here, $g'(t) = \frac{1}{2}J_{\hat{f}}^T(c^{q+1}(t))H_{\hat{f}}^T(c^{q+1}(t))\Delta c^q + \frac{1}{2}(\Delta c^q)^T H_{\hat{f}}(c^{q+1}(t))J_{\hat{f}}^T(c^{q+1}(t))$
 - Both terms are scalars, so $g'(t) = J_{\hat{f}}^T(c^{q+1}(t))H_{\hat{f}}^T(c^{q+1}(t))\Delta c^q$
- Option 3: **Minimize** $\hat{f}(c^{q+1}(t))$ directly (see **unit 16**)

Unit 16

1D Optimization

Part II Roadmap

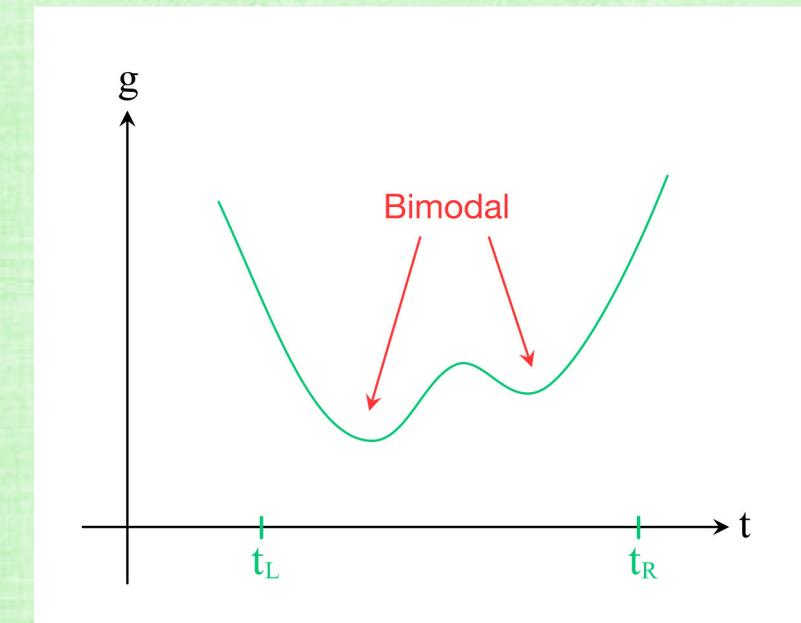
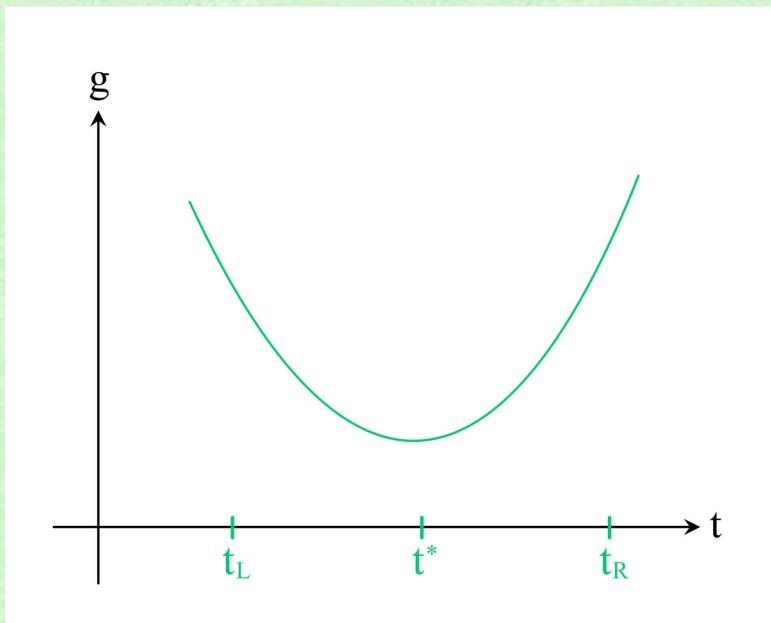
- Part I – Linear Algebra (units 1-12) $Ac = b$
 - Part II – Optimization (units 13-20)
 - (units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima
 - (units 17-18) Computing/Avoiding Derivatives
 - (unit 19) Hack 1.0: “I give up” $H = I$ and J is mostly 0 (descent methods)
 - (unit 20) Hack 2.0: “It’s an ODE!?” (adaptive learning rate and momentum)
-
- ```
graph TD; PartI["Part I – Linear Algebra (units 1-12) $Ac = b$ "] -- "linearize" --> Eq[" $Ac = b$ "]; Eq -- "line search" --> OneD["1D roots/minima"]; subgraph PartII ["Part II – Optimization (units 13-20)"] direction TB; subgraph Theory ["Theory"]; NE1["(units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima"]; NE2["(units 17-18) Computing/Avoiding Derivatives"]; NE3["(unit 19) Hack 1.0: ‘I give up’ $H = I$ and J is mostly 0 (descent methods)"]; NE4["(unit 20) Hack 2.0: ‘It’s an ODE!?’ (adaptive learning rate and momentum)"]; end; subgraph Methods ["Methods"]; end; Theory --> Methods
```

# Leveraging Root Finding (unit 15)

- Relative extrema of  $g(t)$  occur at critical points where  $g'(t) = 0$
- Thus, could directly utilize root finding methods on  $g'$
- Newton:  $t^{q+1} = t^q - \frac{g'(t^q)}{g''(t^q)}$  (dividing by  $g''$  is even worse than dividing by  $g'$ )
- Secant:  $t^{q+1} = t^q - g'(t^q) \frac{t^q - t^{q-1}}{g'(t^q) - g'(t^{q-1})}$  (could replace  $g'$  too)
- Bisection:  $g'(t_L)g'(t_R) < 0$  is the new condition
- Mixed Methods: mixing the above (as per unit 15)

# Unimodal

- Unimodal means one single mode, just as bimodal means two modes
- In 1D optimization, this means that the function has one relative minimum
- $g(t)$  is unimodal in  $[t_L, t_R]$  if and only if  $g$  is monotonically decreasing in  $[t_L, t^*]$  and monotonically increasing in  $[t^*, t_R]$

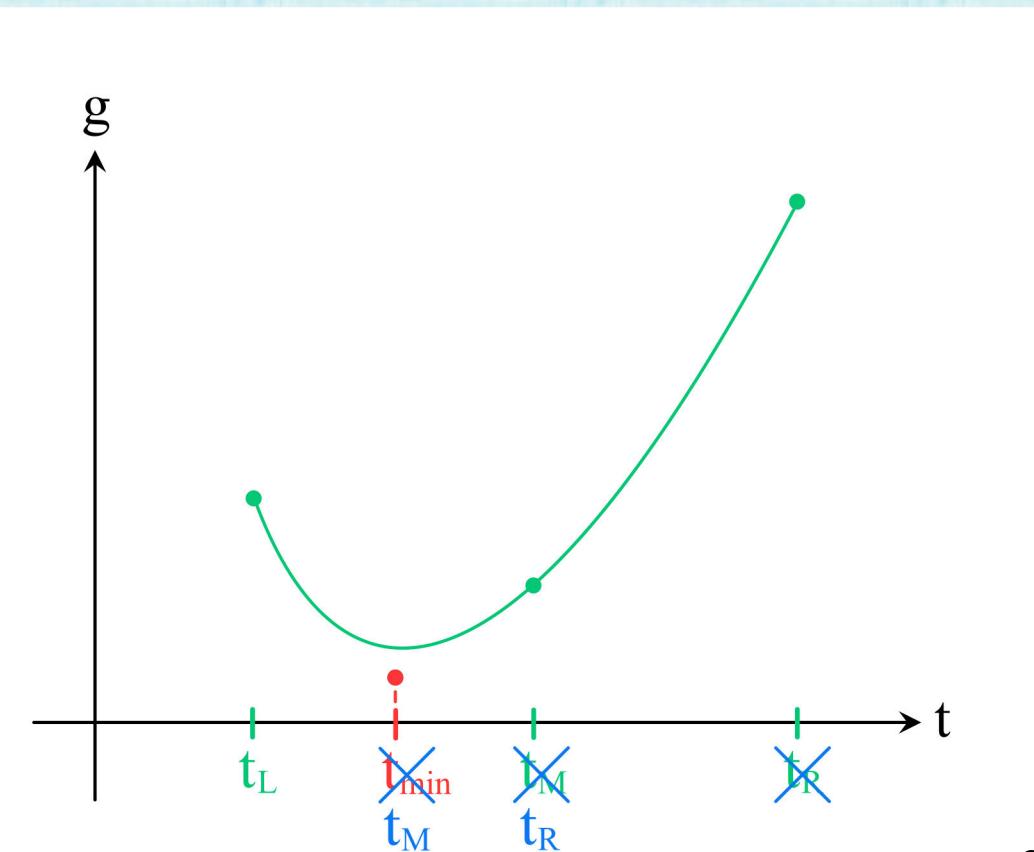
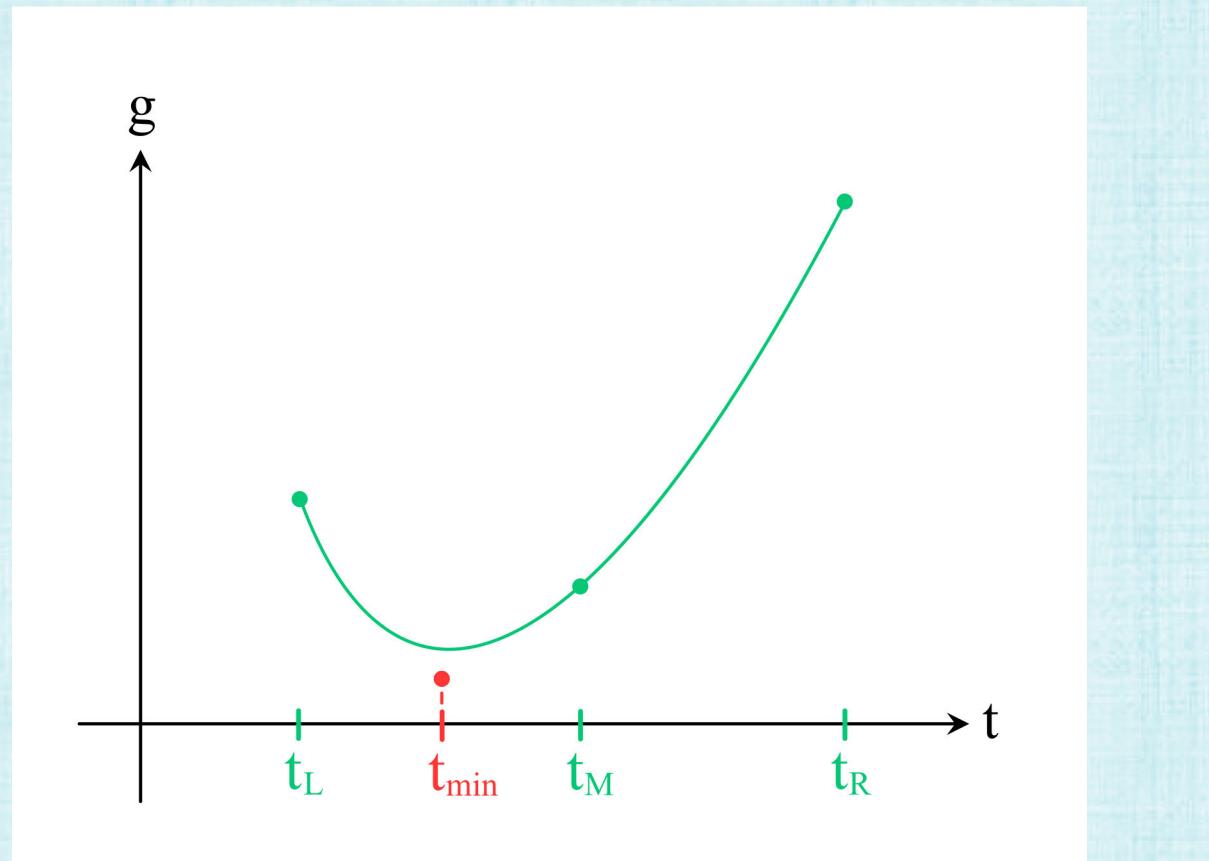


# Successive Parabolic Interpolation

- Motivated by Newton/Secant which use lines to find candidates for roots, instead use parabolas to find candidates for minima
- Given interval  $[t_L, t_R]$  with midpoint  $t_M = \frac{t_L+t_R}{2}$ , create the unique parabola through  $t_L$ ,  $t_R$ , and  $t_M$ 
  - A unimodal  $g$  in  $[t_L, t_R]$  makes this parabola concave up
  - Let  $t_{min}$  be the point where the parabola takes on its minimum value
- Assume  $t_{min} < t_M$  (otherwise, simply rename them)
- If  $g(t_{min}) \leq g(t_M)$ , discard  $[t_M, t_R]$  which cannot contain the minimum
  - Then, set  $t_R = t_M$  and  $t_M = t_{min}$
- If  $g(t_{min}) \geq g(t_M)$ , discard  $[t_L, t_{min}]$  which cannot contain the minimum
  - Then, set  $t_L = t_M$  and  $t_M = t_{min}$
- Superlinear convergence rate with  $p \approx 1.325$

# Successive Parabolic Interpolation

- When  $g(t_{min}) \leq g(t_M)$ , discard  $[t_M, t_R]$  and set  $t_R = t_M$  and  $t_M = t_{min}$



# Golden Section Search

- Unlike bisection for root finding, 3 points is not enough to discard an interval during 1D minimization
- Successive parabolic interpolation demonstrated that 4 points is enough
- Let interval  $[t_L, t_R]$  have intermediate points with  $t_L < t_{M1} < t_{M2} < t_R$ 
  - If  $g$  is unimodal in  $[t_L, t_R]$ , one can safely discard either  $[t_L, t_{M1}]$  or  $[t_{M2}, t_R]$
- If  $g(t_{M1}) \leq g(t_{M2})$ , discard  $[t_{M2}, t_R]$  which cannot contain the minimum
- If  $g(t_{M1}) \geq g(t_{M2})$ , discard  $[t_L, t_{M1}]$  which cannot contain the minimum

# Golden Section Search

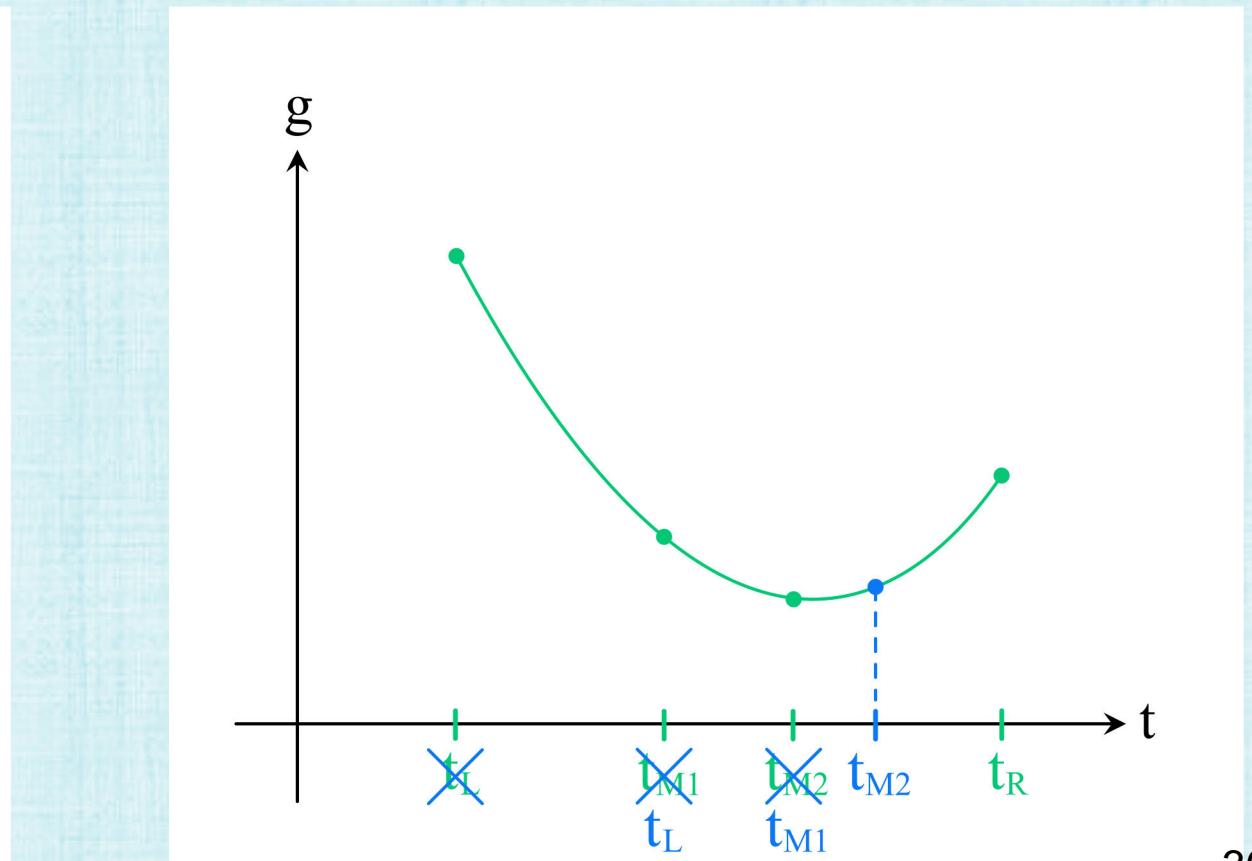
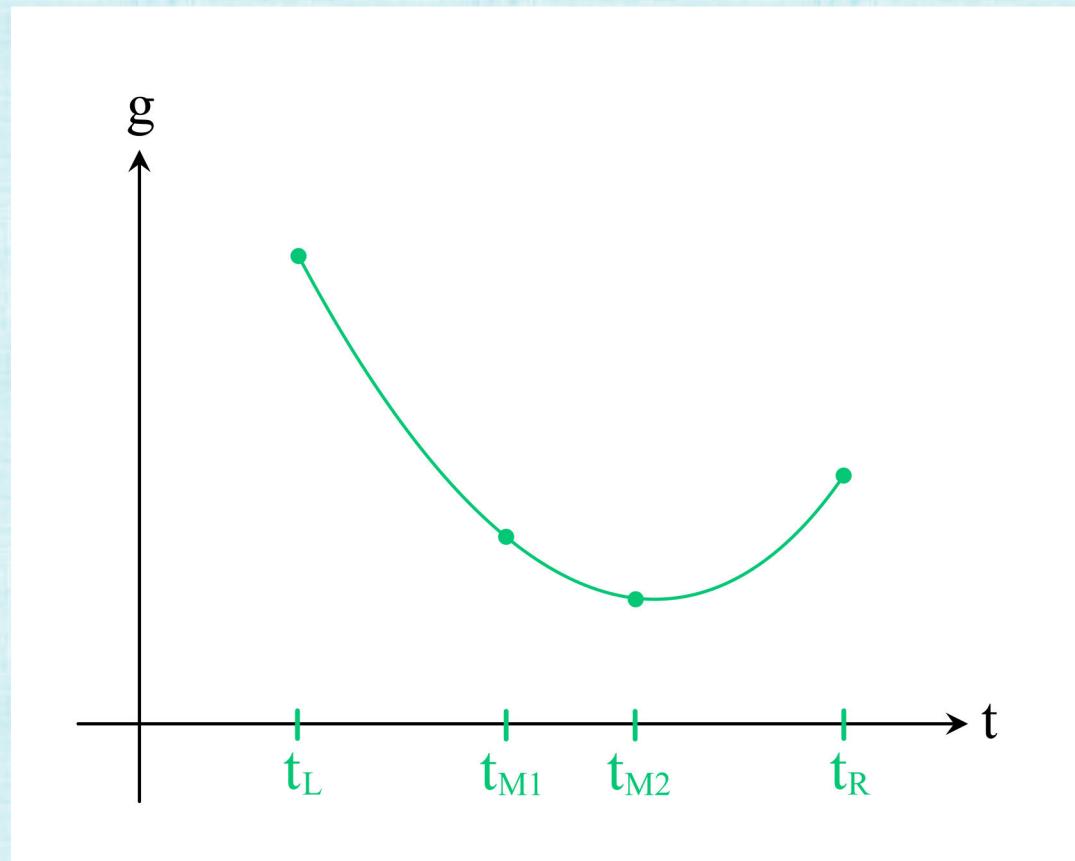
- After discarding an interval, either  $t_{M1}$  or  $t_{M2}$  becomes an endpoint, and keeping the other as an interior point (efficiently) reduces evaluations of  $g$
- Let  $\delta = t_R - t_L$  be the interval size and  $\lambda \in (0, .5)$  be the fraction inward of  $t_{M1}$
- Then  $t_{M1} = t_L + \lambda\delta$ , and symmetric placement gives  $t_{M2} = t_L + \delta - \lambda\delta$
- Discard the left interval (discarding the right gives the same math) to obtain  $t_L^{new} = t_{M1}$  and  $\delta^{new} = (1 - \lambda)\delta$
- Then  $t_{M2} = (t_L^{new} - \lambda\delta) + \delta - \lambda\delta = t_L^{new} + (1 - 2\lambda)\frac{\delta^{new}}{1-\lambda}$  can be designated either  $t_{M1}^{new}$  or  $t_{M2}^{new}$  depending on whether  $\frac{1-2\lambda}{1-\lambda}$  equals  $\lambda$  or  $1 - \lambda$
- Only one of the four solutions has  $\lambda \in (0, .5)$ , i.e.  $t_{M2} = t_{M1}^{new}$  when  $\lambda = \frac{3-\sqrt{5}}{2}$

# Golden Section Search

- Rewriting:  $t_{M1} = (1 - \lambda)t_L + \lambda t_R$  and  $t_{M2} = \lambda t_L + (1 - \lambda)t_R$
- Switch the parameter to the more typical  $\tau = 1 - \lambda = \frac{\sqrt{5}-1}{2}$  (since  $\lambda = \frac{3-\sqrt{5}}{2}$ )
- Then,  $t_{M1} = \tau t_L + (1 - \tau)t_R$  and  $t_{M2} = (1 - \tau)t_L + \tau t_R$
- If  $g(t_{M1}) \leq g(t_{M2})$ , discard  $[t_{M2}, t_R]$ , set  $t_R = t_{M2}$ ,  $t_{M2} = t_{M1}$ , and recompute  $t_{M1}$
- If  $g(t_{M1}) \geq g(t_{M2})$ , discard  $[t_L, t_{M1}]$ , set  $t_L = t_{M1}$ ,  $t_{M1} = t_{M2}$ , and recompute  $t_{M2}$
- Stop when the interval size is small (as usual)
- Linear convergence rate ( $p = 1$ ) with  $C = \frac{1-\lambda}{1} = \tau \approx .618$

# Golden Section Search

- If  $g(t_{M1}) \geq g(t_{M2})$ , discard  $[t_L, t_{M1}]$ , set  $t_L = t_{M1}$ ,  $t_{M1} = t_{M2}$ , recompute  $t_{M2}$



# Mixed Methods

- Given a unimodal interval  $[t_L, t_R]$
- Iterate with Successive Parabolic Interpolation as long as the iterates stay inside the interval
  - When iteration attempts to leave the interval, use prior iterates to shrink the interval as much as possible (while still containing the minima)
- Golden Section Search can be used to continue to shrink the interval, whenever Successive Parabolic Interpolation would fail to stay inside the current interval
- Leverages the speed of Successive Parabolic Interpolation, while still guaranteeing convergence via Golden Section Search
- Many/various strategies exist

# Function/Derivative Requirements

- All methods require function evaluation  $g$
- Root finding approaches differentiate  $g$  and solve  $g'(t) = 0$  to identify critical points
  - All root finding methods require function evaluation, which is  $g'$  here
  - Newton (and mixed methods using Newton) requires the derivative of the function, which is  $g''$  here (since the function is  $g'$ )

# Recall: Useful Derivatives (unit 16)

- $\frac{\partial}{\partial t} c^{q+1}(t) = \Delta c^q$ , since  $c^{q+1}(t) = c^q + t\Delta c^q$
- $\frac{\partial}{\partial t} F(c^{q+1}(t)) = J_F(c^{q+1}(t))\Delta c^q$  and  $\frac{\partial}{\partial t} F^T(c^{q+1}(t)) = (\Delta c^q)^T J_F^T(c^{q+1}(t))$ 
  - $\frac{\partial}{\partial t} F_i(c^{q+1}(t)) = (J_F)_i(c^{q+1}(t)) \Delta c^q$  where  $F_i(c^{q+1}(t))$  are the scalar row components of  $F(c^{q+1}(t))$
- Scalar  $\hat{f}(c^{q+1}(t))$  has system  $J_{\hat{f}}^T(c^{q+1}(t)) = 0$  for critical points
- $\frac{\partial}{\partial t} J_{\hat{f}}^T(c^{q+1}(t)) = H_{\hat{f}}^T(c^{q+1}(t))\Delta c^q$  and  $\frac{\partial}{\partial t} J_{\hat{f}}(c^{q+1}(t)) = (\Delta c^q)^T H_{\hat{f}}(c^{q+1}(t))$ 
  - $\frac{\partial}{\partial t} (J_{\hat{f}}^T)_i(c^{q+1}(t)) = (H_{\hat{f}}^T)_i(c^{q+1}(t))\Delta c^q$

# More Useful Derivatives

- $\frac{\partial}{\partial t} J_F(c^{q+1}(t)) = (\Delta c^q)^T H_F(c^{q+1}(t))$ 
  - $H_F$  is a rank 3 tensor of all 2<sup>nd</sup> derivatives of  $F$
  - $\frac{\partial}{\partial t} (J_F)_i(c^{q+1}(t)) = (\Delta c^q)^T (H_F)_i(c^{q+1}(t))$
- $\frac{\partial}{\partial t} H_{\hat{f}}^T(c^{q+1}(t)) = (\Delta c^q)^T OMG_{\hat{f}}^T(c^{q+1}(t))$ 
  - $OMG_{\hat{f}}^T$  is a rank 3 tensor of all 3<sup>rd</sup> derivatives of  $\hat{f}$
  - $\frac{\partial}{\partial t} \left( H_{\hat{f}}^T \right)_i(c^{q+1}(t)) = (\Delta c^q)^T \left( OMG_{\hat{f}}^T \right)_i(c^{q+1}(t))$

# Nonlinear Systems Problems

- Solve  $J_F(c^q)\Delta c^q = -F(c^q)$  for  $\Delta c^q$  and use  $c^{q+1}(t) = c^q + t\Delta c^q$  in  $F(c^{q+1}(t)) = 0$
- Option 1: For vector valued  $F(c^{q+1}(t))$ , simultaneously (for all  $i$ ) **minimize** all the  $g_i(t) = F_i(c^{q+1}(t))$  aiming for roots where all  $F_i(c^{q+1}(t)) = 0$ 
  - Here,  $g'_i(t) = (J_F)_i(c^{q+1}(t))\Delta c^q$  and  $g''_i(t) = (\Delta c^q)^T(H_F)_i(c^{q+1}(t))\Delta c^q$
- Option 2: **Minimize**  $g(t) = \frac{1}{2}F^T(c^{q+1}(t))F(c^{q+1}(t))$  aiming for its roots
  - Here,  $g'(t) = F^T(c^{q+1}(t))J_F(c^{q+1}(t))\Delta c^q$
  - $g''(t) = F^T(c^{q+1}(t))(\Delta c^q)^TH_F(c^{q+1}(t))\Delta c^q + (\Delta c^q)^TJ_F^T(c^{q+1}(t))J_F(c^{q+1}(t))\Delta c^q$

# Optimization Problems

- Solve  $H_{\hat{f}}^T(c^q)\Delta c^q = -J_{\hat{f}}^T(c^q)$  for  $\Delta c^q$  and use  $c^{q+1}(t) = c^q + t\Delta c^q$  in  $J_{\hat{f}}^T(c^{q+1}(t)) = 0$
- Option 1: For vector valued  $J_{\hat{f}}^T(c^{q+1}(t))$ , simultaneously (for all  $i$ ) **minimize** all the  $g_i(t) = (J_{\hat{f}}^T)_i(c^{q+1}(t))$  aiming for the roots which are critical points of  $\hat{f}(c)$ 
  - Here,  $g'_i(t) = (H_{\hat{f}}^T)_i(c^{q+1}(t))\Delta c^q$  and  $g''_i(t) = (\Delta c^q)^T \left( OMG_{\hat{f}}^T \right)_i(c^{q+1}(t)) \Delta c^q$
- Option 2: **Minimize**  $g(t) = \frac{1}{2} J_{\hat{f}}^T(c^{q+1}(t)) J_{\hat{f}}^T(c^{q+1}(t))$  aiming for the roots which are critical points of  $\hat{f}(c)$ 
  - Here,  $g'(t) = J_{\hat{f}}(c^{q+1}(t)) H_{\hat{f}}^T(c^{q+1}(t)) \Delta c^q$
  - $g''(t) = J_{\hat{f}}(c^{q+1}(t)) (\Delta c^q)^T OMG_{\hat{f}}^T(c^{q+1}(t)) \Delta c^q + (\Delta c^q)^T H_{\hat{f}}(c^{q+1}(t)) H_{\hat{f}}^T(c^{q+1}(t)) \Delta c^q$
- Option 3: **Minimize**  $\hat{f}(c^{q+1}(t))$  directly
  - $g'(t) = J_{\hat{f}}(c^{q+1}(t)) \Delta c^q$  and  $g''(t) = (\Delta c^q)^T H_{\hat{f}}(c^{q+1}(t)) \Delta c^q$

# Unit 17

# Computing Derivatives

# Part II Roadmap

- Part I – Linear Algebra (units 1-12)  $Ac = b$ 
    - Part II – Optimization (units 13-20)
      - (units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima
      - (units 17-18) Computing/Avoiding Derivatives
      - (unit 19) Hack 1.0: “I give up”  $H = I$  and  $J$  is mostly 0 (descent methods)
      - (unit 20) Hack 2.0: “It’s an ODE!?” (adaptive learning rate and momentum)
- 
- ```
graph TD; A["Part I – Linear Algebra (units 1-12)  $Ac = b$ "] -- "linearize" --> B[" $Ac = b$ "]; B -- "line search" --> C["Part II – Optimization (units 13-20)"]; C -- "Theory" --> D["(units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima"]; C -- "Methods" --> E["(units 17-18) Computing/Avoiding Derivatives"]; C -- "Methods" --> F["(unit 19) Hack 1.0: ‘I give up’  $H = I$  and  $J$  is mostly 0 (descent methods)"]; C -- "Methods" --> G["(unit 20) Hack 2.0: ‘It’s an ODE!?’ (adaptive learning rate and momentum)"];
```

Smoothness

- Discontinuous functions cannot be differentiated
 - Even methods that don't require derivatives struggle when functions are discontinuous
- Continuous functions may have kinks (discontinuities in derivatives)
 - Discontinuous derivatives cause methods that depend on derivatives to fail, since function behavior cannot be adequately predicted from one side of the kink to the other
- Typically, functions need to be “smooth enough”, which has varying meaning depending on the approach
- Specialty approaches for special classes of functions, e.g. linear algebra, linear programming, convex optimization, second order cone program, etc.
 - Nonlinear Systems/Optimization are so difficult that they often receive less scrutiny/care, as best practices/techniques often do not exist

Biological Neurons (towards real AI)

- Aim to mimic human biological neural networks and learning
- Biological neurons are “all or none”, which motivates similar strategies in artificial neural networks
 - This leads to a discontinuous function with an identically zero derivative everywhere else
 - Disastrous for optimization!
- Biological neurons fire with increased frequency for stronger signals
 - This leads to a piecewise constant and discontinuous derivative
 - Problematic for optimization!
- Smoothing allows optimization to work, i.e. to minimize the loss to find the parameters/coefficients for the network architecture

Heaviside Function

- $H(x) = 1$ for $x \geq 0$, and $H(x) = 0$ for $x < 0$
- Motivated by biological neurons being “all or none”, but has a discontinuity at 0 and derivative identically zero elsewhere

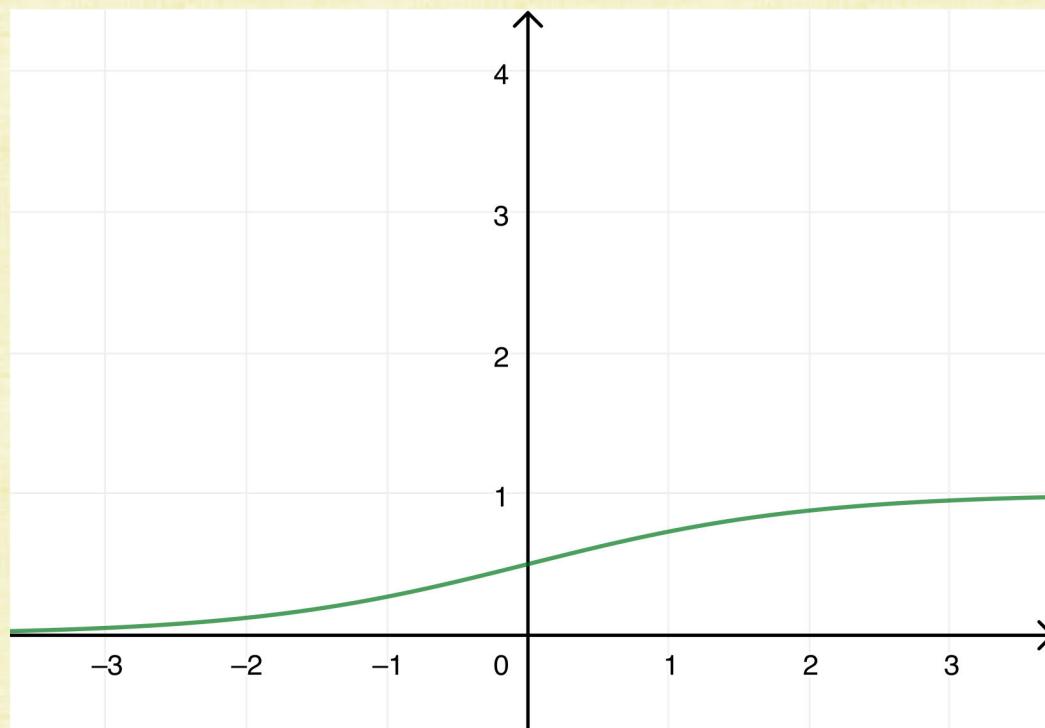


(Inequality) Constrained Optimization

- Minimize $\hat{f}(c)$ subject to $\hat{g}(c) \geq 0$ (or strictly $\hat{g}(c) > 0$)
- Heaviside function can be used to create a penalty term $-H(-\hat{g}_i(c))\hat{g}_i(c)$ which is only nonzero when $\hat{g}_i(c) < 0$
 - This penalty term is minimized by forcing negative $\hat{g}_i(c)$ towards zero (as desired)
- Given diagonal matrix D of (positive) weights indicating the relative importance of various constraints, unconstrained optimization can be used to minimize
$$\hat{f}(c) - \sum_i H(-\hat{e}_i^T D \hat{g}(c)) \hat{e}_i^T D \hat{g}(c)$$
 - However, this requires differentiating the non-smooth Heaviside function
 - Smoothing the Heaviside function makes the cost function differentiable

Sigmoid Function (an example)

- Any smoothed Heaviside function, e.g. $S(x) = \frac{1}{1+e^{-x}}$ (there are many options)
- Continuous and monotonically increasing, although the derivative is close to zero away from $x = 0$



Example: Binary Classification

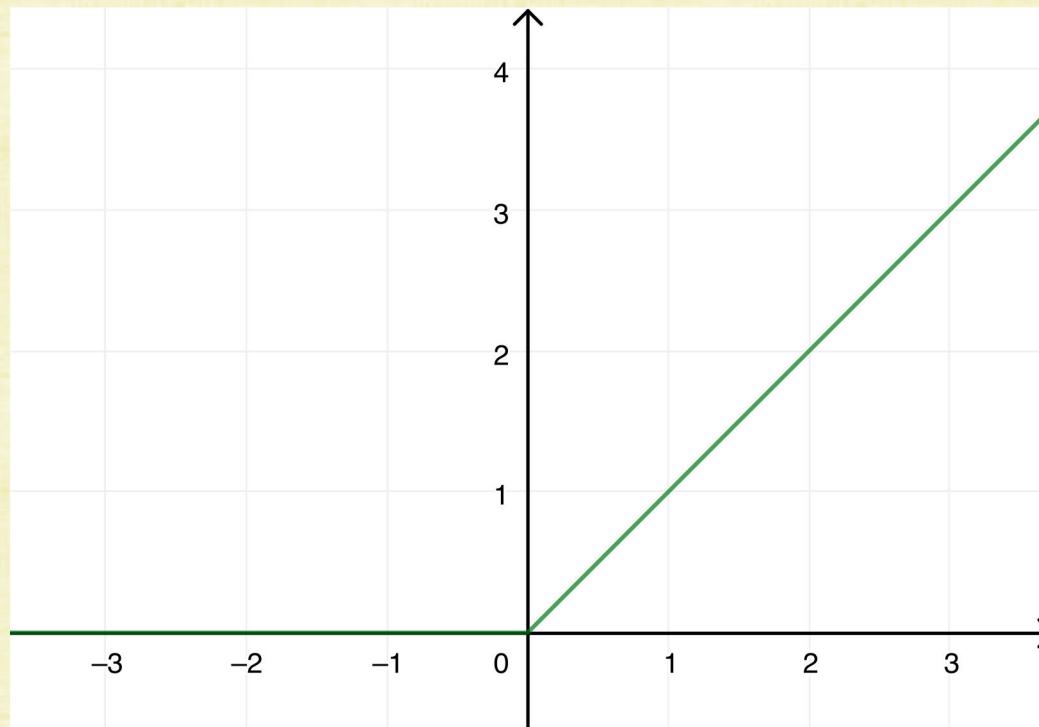
- Training data (x_i, y_i) where the $y_i = \pm 1$ are binary class labels
- Find hyperplane $n^T(x - x_o) = 0$ that separates the data between the two class labels (n is the unit normal and x_o is a point on the plane)
- The closest x_i on each side of this hyperplane are called the support vectors
- If the hyperplane is equidistant between the support vectors, then they lie on parallel planes: $n^T(x - x_o) = \pm\epsilon$ (where ϵ is the margin)
- Dividing by ϵ to normalize gives $c^T(x - x_o) = \pm 1$ where c is in the normal direction (but not unit length), and then maximizing the margin ϵ is equivalent to minimizing $\|c\|_2$
- That is, minimize $\hat{f}(c) = \frac{1}{2}c^T c$ while still fitting the data

Example: Binary Classification

- Minimize $\hat{f}(c) = \frac{1}{2} c^T c$ subject to inequality constraints
- $c^T(x_i - x_o) \geq 1$ when $y_i = 1$ and $c^T(x_i - x_o) \leq -1$ when $y_i = -1$ can be combined into $y_i c^T(x_i - x_o) \geq 1$ for every data point
- Alternatively, $y_i(c^T x_i - b) \geq 1$ where the new scalar unknown is $b = c^T x_o$
- New data will be inferred/classified based on the sign of $c^T x_{new} - b$
- When approached via unconstrained optimization, the Heaviside function incorporates constraints into the cost function (and subsequently smoothing the Heaviside is called soft-margin)

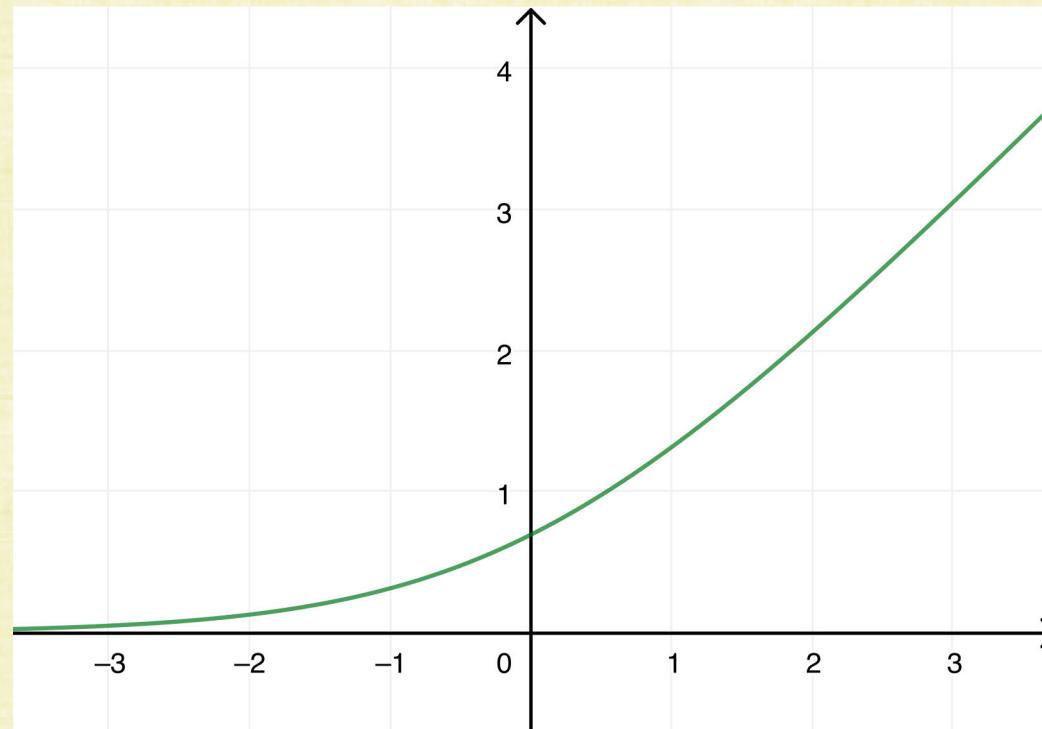
Rectifier Functions (an example)

- $R(x) = \max(x, 0)$ or similar functions which are continuous and have increasing values
- Motivated by biological neurons firing with increased frequency for stronger signals
- Piecewise constant and discontinuous derivative causes issues with optimization



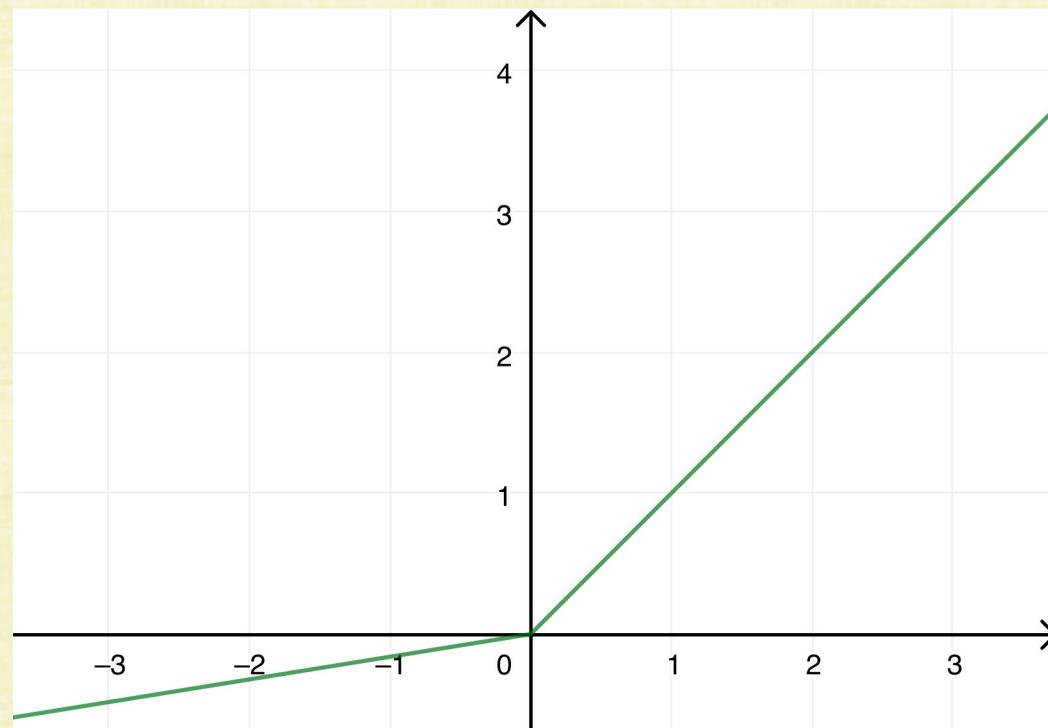
Softplus Function

- Softplus function $SP(x) = \log(1 + e^x)$ smooths the discontinuous derivative typical of rectifier functions



Leaky Rectifier Function

- Modifies the negative part of a rectifier function to also have a positive slope instead of being set to zero
- Can be smoothed (as well)



Arg/Soft Max

- Arg Max returns 1 for the largest argument and 0 for the other arguments
- E.g. (.99,1) → (0,1), (1,.99) → (1,0), etc.
- Highly discontinuous!
- Soft Max is a smoothed out version, e.g. $(x_1, x_2) \rightarrow \left(\frac{e^{x_1}}{e^{x_1} + e^{x_2}}, \frac{e^{x_2}}{e^{x_1} + e^{x_2}} \right)$
- This is a smooth function of the arguments, differentiable, etc.
- Variants/weightings exist to make it closer/further from Arg Max (while preserving differentiability)

Symbolic Differentiation

- When a function is known in closed form, it can be differentiated by hand
- Software packages such as Mathematica can aid in symbolic differentiation (and subsequent simplification)
- Some benefits of knowing the closed form derivative:
 - Provides a better understanding of the underlying problem
 - Enables well thought out smoothing/regularization
 - Allows one to implement highly efficient code
 - Subsequently allows access to more accurate higher derivatives
 - Some of the aforementioned benefits enable the use of better solvers
 - Helps to write/maintain code with less bugs
 - Etc.

Symbolic Differentiation of Code

- Sometimes a function is not analytically known and/or merely represents the output of some source code
- Often parts of code have known derivatives, and those known derivatives can be utilized/leveraged via the mathematical rules for differentiation
- Moreover, when parts of the code are always used consecutively, they can be merged; subsequently, merged code with known derivatives in each part can often have the derivative treatment simplified for robustness/efficiency

Differentiate the Right Thing

- Consider an iterative solver, e.g. CG, that solves $Ac = b$ to find c given b
- Furthermore, suppose that the code is enormous, complicated, confusing, a black box, etc. (basically impenetrable)
- It is tempting to consider some of the code bases that claim to differentiate such chunks of code
 - In fact, many times these approaches work, and the answers are reasonable
 - Though it is always hard to know whether computational inaccuracies (as discussed in this class) are having an adverse effect in this black box approach
- On the other hand, when invertible, $c = A^{-1}b$ and $\frac{\partial c_k}{\partial b_i} = \tilde{a}_{ik}$ where \tilde{a}_{ik} is an entry in A^{-1} (a similar approach can be taken for A^+)
- That is, the derivative is independent of the iterative solver and the errors that might accumulate within it due to poor conditioning

Used Car Salesmen

- Beware of the claim that it is good to be able to use something without understanding it
- The claim is true, and many of us enjoy driving our cars without knowledge or care of what is under the hood
- However, those who design cars, manufacture cars, repair cars, etc. benefit greatly from understanding as much as possible about them, and we too benefit enormously from their expertise
- Though, admittedly, there are those in the car business, such as used car salesmen, that authentically do not require any real knowledge/expertise
- The question is: **what kind of computer scientist do you want to be?**

Oversimplified Thinking

- Beware of claims that drastically oversimplify
- E.g., some say that code is very simple and merely consists of simple operations like add/subtract/multiply/divide that are easily differentiated
- However, in reality, even the simple $z = x + y$ has subtleties that can matter
- The computer actually executes $z = \text{round}(x + y)$
- Too many claim that issues they have not carefully considered don't matter in practice; meanwhile, many of the practices are not well understood (leaving one to question the first claim)

Finite Differences

- Derivatives can be approximated by various formulas, which (recall) is how the Secant method was derived from Newton's method
- Given a small perturbation h , Taylor expansions can be manipulated to write:
 - Forward Difference: $g'(t) = \frac{g(t+h)-g(t)}{h} + O(h)$, 1st order accurate
 - Backward Difference: $g'(t) = \frac{g(t)-g(t-h)}{h} + O(h)$, 1st order accurate
 - Central Difference: $g'(t) = \frac{g(t+h)-g(t-h)}{2h} + O(h^2)$, 2nd order accurate
 - Second Derivative: $g''(t) = \frac{g(t+h)-2g(t)+g(t-h)}{h^2} + O(h^2)$, 2nd order accurate
- These approximations can be evaluated even when $g(t)$ is not known precisely but merely represents the output of some code with input t

Finite Differences (Drawbacks)

- Finite Differences only give an approximation to the derivative, and contain truncation errors related to the perturbation size h
- One has to reason about the effects that truncation error and the size of h have on other aspects of the code
- If the code is very long and complex, the overall effects of truncation errors may be unclear
- Still, finite difference methods have had a broad positive impact in computational science!

Automatic Differentiation

- In machine learning, this is often referred to as Back Propagation
- For every (potentially vector valued) function $F(c_{input})$ written into the code, an analytically correct companion function for the Jacobian matrix $\frac{\partial F}{\partial c}(c_{input})$ is also written into the code
- Then when evaluating $F(c_{input})$, one can also evaluate $\frac{\partial F}{\partial c}(c_{input})$
 - Of course, $\frac{\partial F}{\partial c}(c_{input})$ contains roundoff errors based on machine precision (and conditioning, etc.)
 - But it does not contain the much larger truncation errors present in finite differencing
- Code chunks combine together various functions via arithmetic/compositional rules
- Analytic differentiation has its own set of rules (linearity, product rule, quotient rule, chain rule, etc.) that are used to assemble the derivative (evaluated at c_{input}) for the code chunk
 - Roundoff errors will accumulate, of course, and the resulting error has the potential to be catastrophic
 - Similar (potentially worse) sentiments hold for the much larger truncation errors

Second Derivatives

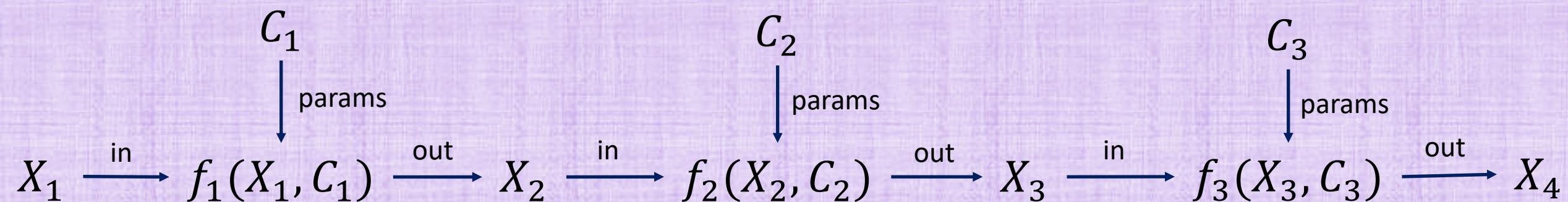
- If c_{input} is size n and $F(c_{input})$ is size m , then the Jacobian matrix $\frac{\partial F}{\partial c}(c_{input})$ is size $m \times n$
- The Hessian of second derivatives is a size $m \times n \times n$
 - Note that $m = 1$ for optimization, i.e. for $\hat{f}(c_{input})$
- Writing automatic differentiating functions for all second derivatives can be difficult/tedious
- Storing Hessians for all second derivatives can be unwieldy/intractable
- Roundoff error accumulation is an even bigger problem for second derivatives, and the resulting errors are even more likely to have adverse effects
- Additional smoothness is required for second derivatives as well
 - This is a problem for any method that considers second derivatives, and is not specific to automatic differentiation approaches

Dropout

- One way to combat overfitting is to train several different network architectures on the same data, inference them all, and average the result
 - This is costly, especially if there are many networks
- Dropout is a “hacky” approach to achieving a network function averaged over multiple network architectures
 - Though Google did patent it!
- The idea is to simply ignore parts of the code with some probability when training the network, mimicking a perturbed network architecture
- Although this can be seen as computing correct derivatives on perturbed functions, it can also equivalently be seen as adding uncertainty to the derivative computation
- That is, instead of regularization via model averaging, it can be seen as creating a network robust to errors in derivative estimation

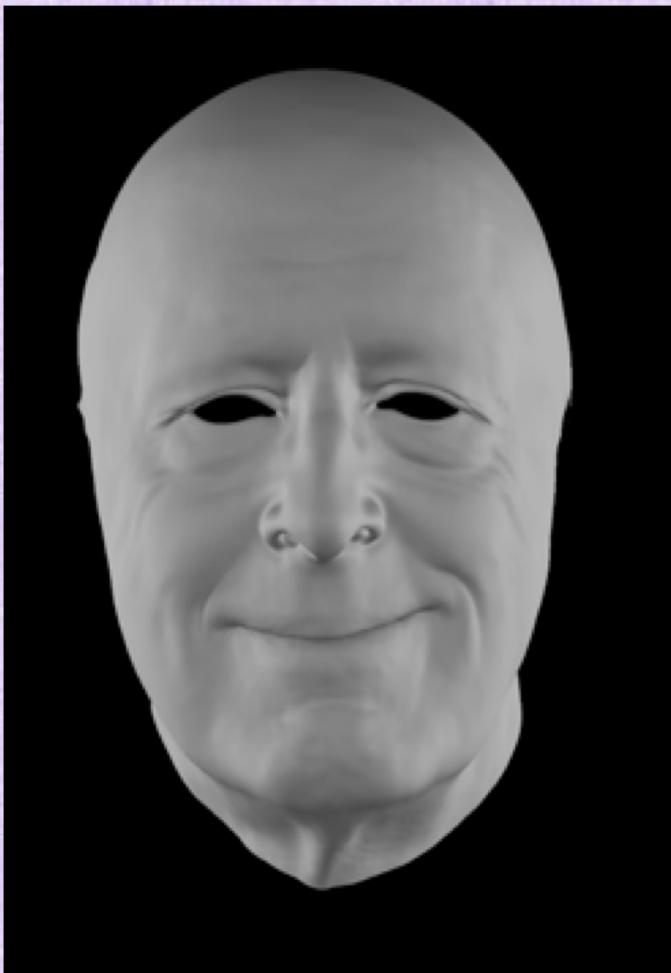
Function Layers

- More complex processes work in a pipeline with many complex layers
- Each layer completes a tasks on its inputs X_j to create outputs X_{j+1}
- Each layer may depend on parameters C_j
- There may be a known/desired output X_{target} to compare the final result to



$$\hat{f}(X_4) = \|X_4 - X_{target}\|$$

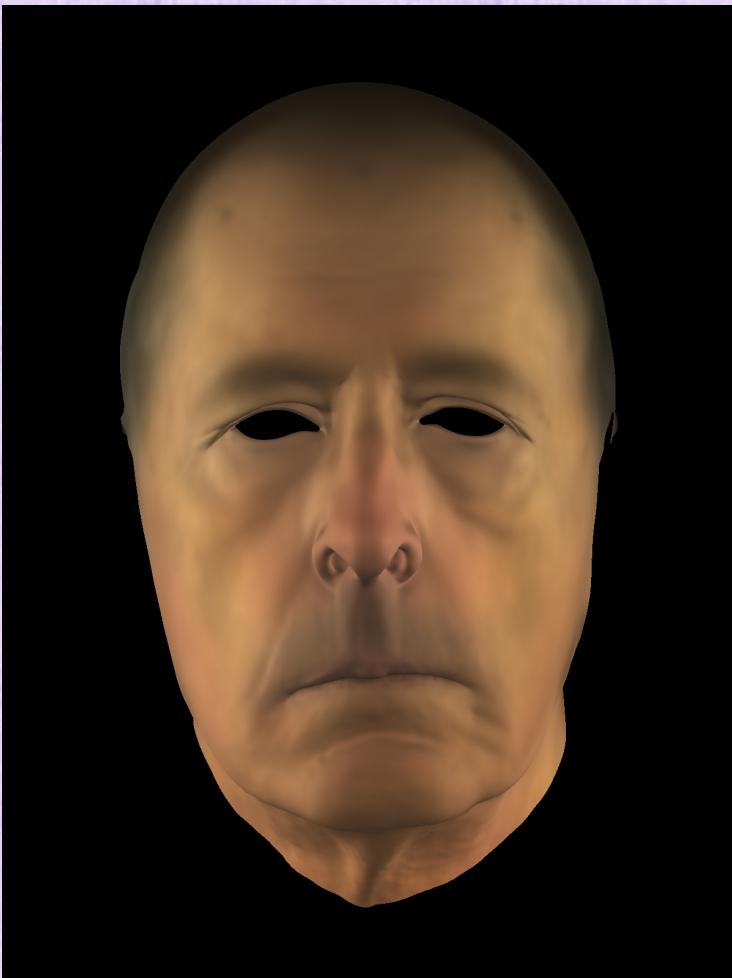
Function Layers (Example)



LAYER 1

- Input: animation controls
- Function: linear blend shapes, nonlinear skinning, quasistatic physics simulation, etc. to deform a face
- Parameters: lots of hand tuned or known parameters including shape libraries, etc.
- Output: vertex positions of a triangle mesh

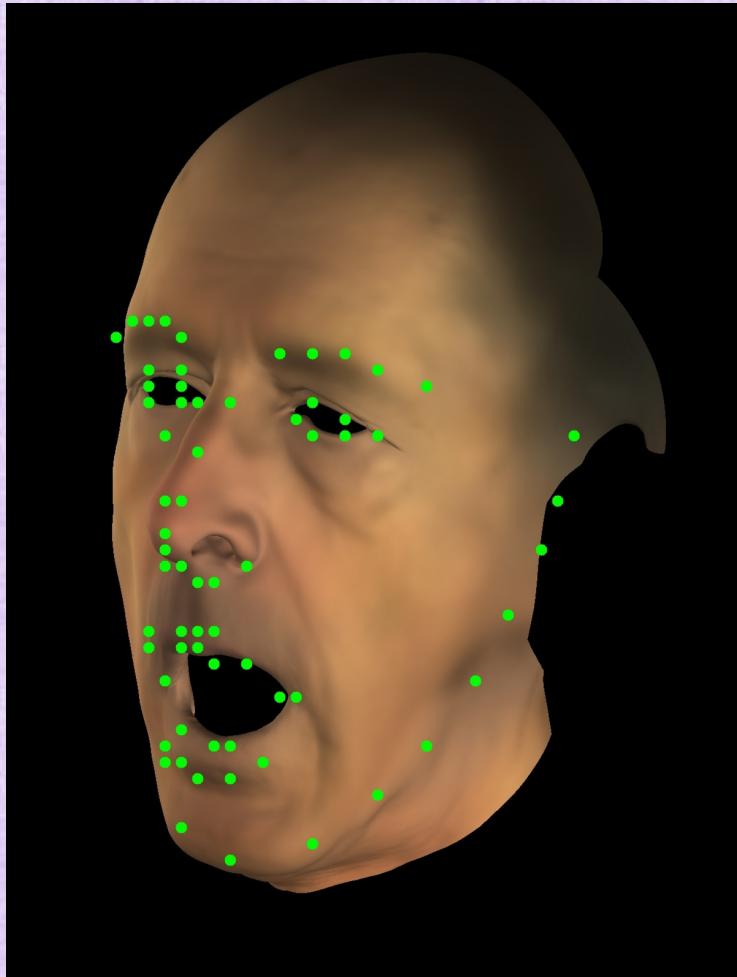
Function Layers (Example)



LAYER 2

- Input: vertex positions of a triangle mesh
- Function: scanline renderer or ray tracer
- Parameters: lots of hand tuned or known parameters for material models, lighting and shading, textures, etc.
- Output: RGB colors for pixels (an image)

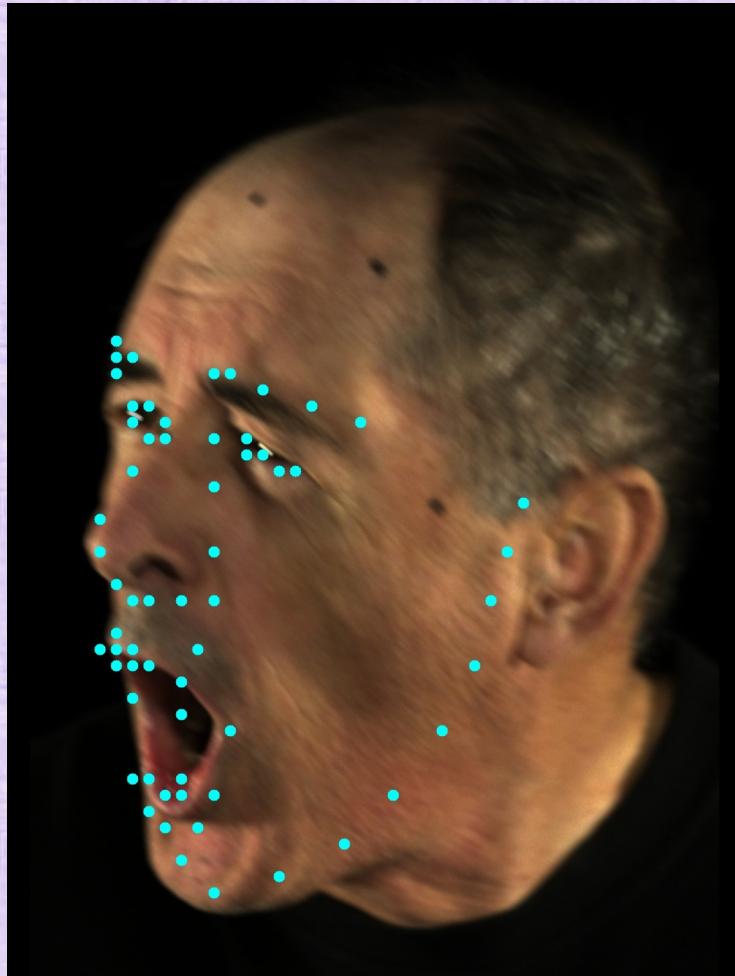
Function Layers (Example)



LAYER 3

- Input: RGB colors for pixels (an image)
- Function: facial landmark detector
- Parameters: parameters for the network architecture determined by training the network to match hand labeled data
- Output: 2D locations of landmarks on the image

Function Layers (Example)

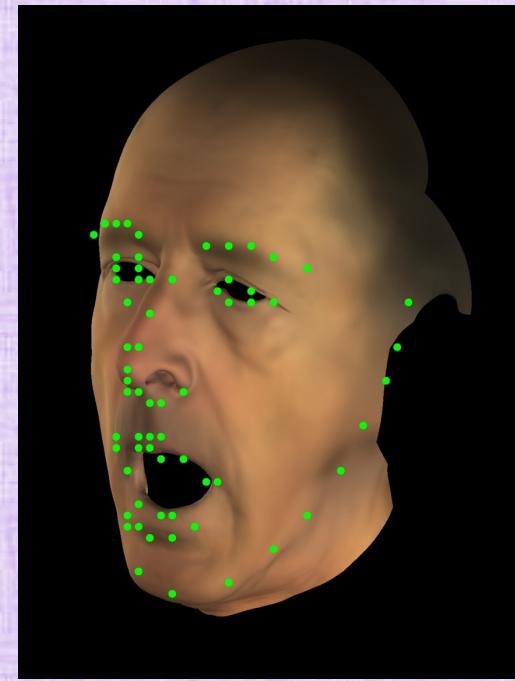
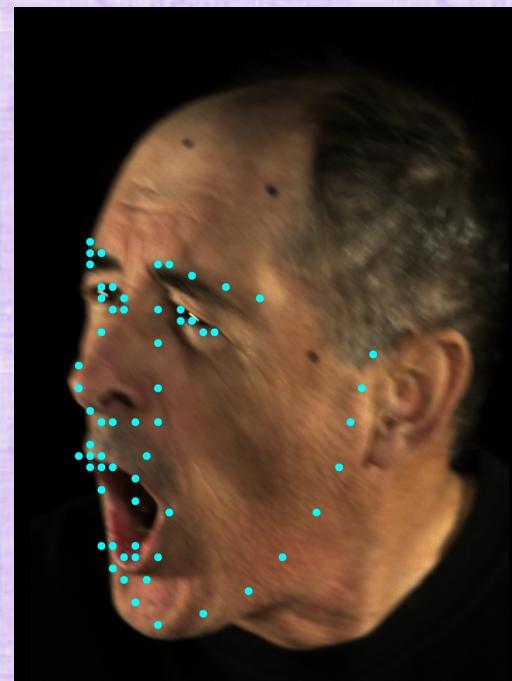


TARGET OUTPUT

- Run a landmark detector on a photograph of the individual to obtain 2D landmark positions
- The goal is to have the 2D landmarks output from the complex multi-layered function match the 2D landmarks on the photograph

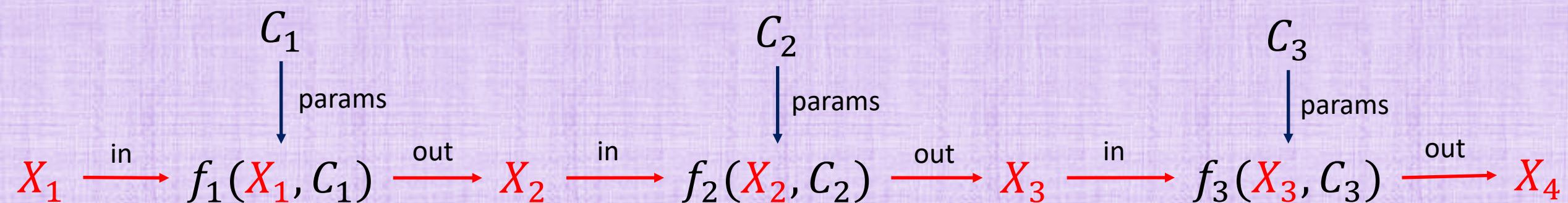
Function Layers (Example)

- Modifying animation controls changes the triangulated surface which changes the rendered pixels in the image which changes the network's determination of landmarks
- When the two sets of landmarks agree, the animation controls indicate what the person in the photograph was doing



Classical Optimization

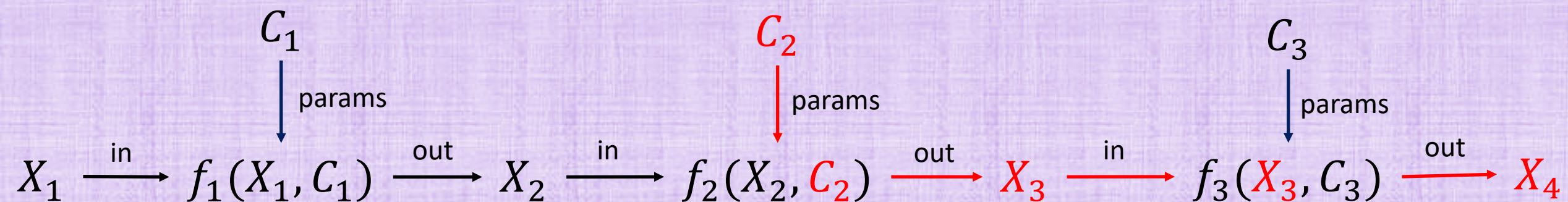
- Find the input X_1 that minimizes $\hat{f}(X_4)$
- Chain rule: $\frac{\partial \hat{f}(X_4)}{\partial X_1} = \frac{\partial \hat{f}(X_4)}{\partial X_4} \frac{\partial X_4}{\partial X_3} \frac{\partial X_3}{\partial X_2} \frac{\partial X_2}{\partial X_1} = \frac{\partial \hat{f}(X_4)}{\partial X_4} \frac{\partial f_3(X_3, C_3)}{\partial X_3} \frac{\partial f_2(X_2, C_2)}{\partial X_2} \frac{\partial f_1(X_1, C_1)}{\partial X_1}$
- Parameters are considered fixed/constant



$$\hat{f}(X_4) = \|X_4 - X_{target}\|$$

Network Training

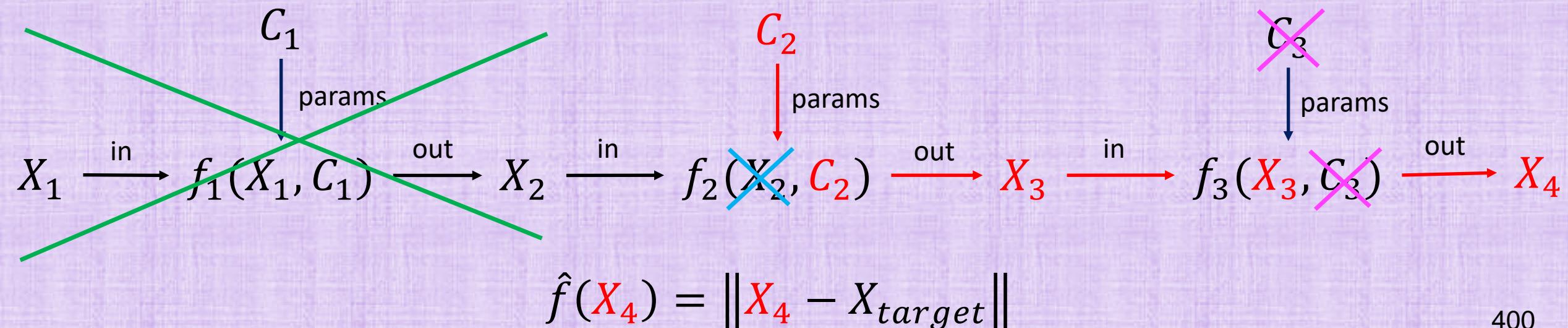
- Train network f_2 by finding parameters C_2 that minimize $\hat{f}(X_4)$
- Chain rule: $\frac{\partial \hat{f}(X_4)}{\partial C_2} = \frac{\partial \hat{f}(X_4)}{\partial X_4} \frac{\partial X_4}{\partial X_3} \frac{\partial X_3}{\partial C_2} = \frac{\partial \hat{f}(X_4)}{\partial X_4} \frac{\partial f_3(X_3, C_3)}{\partial X_3} \frac{\partial f_2(X_2, C_2)}{\partial C_2}$



$$\hat{f}(X_4) = \|X_4 - X_{target}\|$$

Network Training

- Any preprocess to the network does not require differentiability
- The network itself only requires differentiability in terms of its parameters
- Any postprocess to the network requires input/output differentiability, but does not require differentiability in terms of its parameters



Unit 18

Avoiding Derivatives

Part II Roadmap

- Part I – Linear Algebra (units 1-12) $Ac = b$
 - Part II – Optimization (units 13-20)
 - (units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima
 - (units 17-18) Computing/**Avoiding Derivatives**
 - (unit 19) **Hack 1.0:** “I give up” $H = I$ and J is mostly 0 (descent methods)
 - (unit 20) **Hack 2.0:** “It’s an ODE!?” (adaptive learning rate and momentum)
-
- linearize
- line search
- Theory
- Methods

1D Root Finding (Unit 15)

- Newton's method requires g' , as do mixed methods using Newton
- Secant method replaces g' with a secant line though two prior iterates
- Finite differencing (unit 17) may be used to approximate this derivative as well, although one needs to determine the size of the perturbation h
- Automatic differentiation (unit 17) may be used to find the value of g' at a particular point, if/when “backprop” code exists, even when g and g' are not known in closed form
- Convergence is only guaranteed under certain conditions, emphasizing the importance of safe methods (such as mixed methods with bisection)
- Safe methods help to guard against errors in derivative approximations

1D Optimization (Unit 16)

- Root finding approaches search for critical points as the roots of g'
 - All root finding methods use the function itself (g' here)
 - Newton (and mixed methods using Newton) require the derivative of the function (g'' here)
- Can use secant lines for g' and interpolating parabolas for g'' , using either prior iterates (unit 16) or finite differences (unit 17)
- Automatic differentiation (unit 17) may be leveraged as well
 - Although, not (typically) for approaches that require g''
- Safe methods (such as mixed methods with bisection or golden section search) help to guard against errors in the approximation of various derivatives

Nonlinear Systems (Unit 14)

- $J_F(c^q)\Delta c^q = -F(c^q)$ is solved to find the search direction Δc^q
 - Then line search utilizes various 1D approaches (unit 15/16)
- The Jacobian matrix of first derivatives $J_F(c^q)$ needs to be evaluated (given c^q)
- Each entry $\frac{\partial F_i}{\partial c_k}(c^q)$ can be approximated via finite differences (unit 17) or automatic differentiation (unit 17)
- Quasi-Newton approaches get quite cavalier with the idea of a search direction, and as such make various aggressive approximations to the Jacobian $J_F(c^q)$
- Quasi-Newton can wildly perturb the search direction, so **robust/safe approaches to the 1D line search become quite important to making “progress” towards solutions**

Broyden's Method

- An initial guess for the Jacobian is continuously corrected with rank one updates, similar in spirit to a secant approach
- Let $J^0 = I$
- Solve $J^q \Delta c^q = -F(c^q)$ to find search direction Δc^q
 - Use line search to find c^{q+1} and $F(c^{q+1})$, and then update $\Delta c^q = c^{q+1} - c^q$
- Update $J^{q+1} = J^q + \frac{1}{(\Delta c^q)^T \Delta c^q} (F(c^{q+1}) - F(c^q) - J^q \Delta c^q)(\Delta c^q)^T$
- Note: $J^{q+1}(c^{q+1} - c^q) = F(c^{q+1}) - F(c^q)$
- That is, J^{q+1} satisfies a secant type equation $J \Delta c = \Delta F$

Optimization (Unit 13)

- Scalar cost function $\hat{f}(c)$ has critical points where $J_{\hat{f}}^T(c) = 0$ (unit 13)
- $H_{\hat{f}}^T(c^q)\Delta c^q = -J_{\hat{f}}^T(c^q)$ is solved to find a search direction Δc^q (unit 14)
- Then line search utilizes various 1D approaches (unit 15/16)
- The Hessian matrix of second derivatives $H_{\hat{f}}^T(c^q)$ and the Jacobian vector of first derivatives $J_{\hat{f}}^T(c^q)$ both need to be evaluated (given c^q)
- The various entries can be evaluated via finite differences (unit 17) or automatic differentiation (unit 17)
- These approaches can struggle on the Hessian matrix of second partial derivatives
- This makes Quasi-Newton approaches quite popular for optimization
 - When c is large, the $O(n^2)$ Hessian H is unwieldy/intractable, so some approaches approximate the action of H^{-T} on a vector (i.e., on the right hand side)

Broyden's Method (for Optimization)

- Same formulation as for nonlinear systems
- Solve for the search direction, and find c^{q+1} and $J_{\hat{f}}^T(c^{q+1})$
- Update $\Delta c^q = c^{q+1} - c^q$ and $\Delta J_{\hat{f}}^T = J_{\hat{f}}^T(c^{q+1}) - J_{\hat{f}}^T(c^q)$
- Then $(H_{\hat{f}}^T)^{q+1} = (H_{\hat{f}}^T)^q + \frac{1}{(\Delta c^q)^T \Delta c^q} \left(\Delta J_{\hat{f}}^T - (H_{\hat{f}}^T)^q \Delta c^q \right) (\Delta c^q)^T$
- So that $(H_{\hat{f}}^T)^{q+1} \Delta c^q = \Delta J_{\hat{f}}^T$

Broyden's Method (for Optimization)

- For the inverse, using $\Delta c^q = c^{q+1} - c^q$ and $\Delta J_{\hat{f}}^T = J_{\hat{f}}^T(c^{q+1}) - J_{\hat{f}}^T(c^q)$
- Define $(H_{\hat{f}}^{-T})^{q+1} = (H_{\hat{f}}^{-T})^q + \frac{(\Delta c^q - (H_{\hat{f}}^{-T})^q \Delta J_{\hat{f}}^T) (\Delta c^q)^T (H_{\hat{f}}^{-T})^q}{(\Delta c^q)^T (H_{\hat{f}}^{-T})^q \Delta J_{\hat{f}}^T}$
- So that $(H_{\hat{f}}^{-T})^{q+1} \Delta J_{\hat{f}}^T = \Delta c^q$
- Then, solving $H_{\hat{f}}^T(c^{q+1}) \Delta c^{q+1} = -J_{\hat{f}}^T(c^{q+1})$ is replaced with defining the search direction by $\Delta c^{q+1} = -(H_{\hat{f}}^{-T})^{q+1} J_{\hat{f}}^T(c^{q+1})$

SR1 (Symmetric Rank 1)

- For the inverse, using $\Delta c^q = c^{q+1} - c^q$ and $\Delta J_{\hat{f}}^T = J_{\hat{f}}^T(c^{q+1}) - J_{\hat{f}}^T(c^q)$
- Define $(H_{\hat{f}}^{-T})^{q+1} = (H_{\hat{f}}^{-T})^q + \frac{(\Delta c^q - (H_{\hat{f}}^{-T})^q \Delta J_{\hat{f}}^T) (\Delta c^q - (H_{\hat{f}}^{-T})^q \Delta J_{\hat{f}}^T)^T}{(\Delta c^q - (H_{\hat{f}}^{-T})^q \Delta J_{\hat{f}}^T)^T \Delta J_{\hat{f}}^T}$
- So that $(H_{\hat{f}}^{-T})^{q+1} \Delta J_{\hat{f}}^T = \Delta c^q$
- Then, solving $H_{\hat{f}}^T(c^{q+1}) \Delta c^{q+1} = -J_{\hat{f}}^T(c^{q+1})$ is replaced with defining the search direction by $\Delta c^{q+1} = -(H_{\hat{f}}^{-T})^{q+1} J_{\hat{f}}^T(c^{q+1})$

DFP (Davidon-Fletcher-Powell)

- For the inverse, using $\Delta c^q = c^{q+1} - c^q$ and $\Delta J_{\hat{f}}^T = J_{\hat{f}}^T(c^{q+1}) - J_{\hat{f}}^T(c^q)$
- Define $(H_{\hat{f}}^{-T})^{q+1} = (H_{\hat{f}}^{-T})^q - \frac{(H_{\hat{f}}^{-T})^q \Delta J_{\hat{f}}^T \Delta J_{\hat{f}} (H_{\hat{f}}^{-T})^q}{\Delta J_{\hat{f}} (H_{\hat{f}}^{-T})^q \Delta J_{\hat{f}}^T} + \frac{\Delta c^q (\Delta c^q)^T}{(\Delta c^q)^T \Delta J_{\hat{f}}^T}$
- So that $(H_{\hat{f}}^{-T})^{q+1} \Delta J_{\hat{f}}^T = \Delta c^q$
- Then, solving $H_{\hat{f}}^T(c^{q+1}) \Delta c^{q+1} = -J_{\hat{f}}^T(c^{q+1})$ is replaced with defining the search direction by $\Delta c^{q+1} = -(H_{\hat{f}}^{-T})^{q+1} J_{\hat{f}}^T(c^{q+1})$

BFGS (Broyden-Fletcher-Goldfarb-Shanno)

- For the inverse, using $\Delta c^q = c^{q+1} - c^q$ and $\Delta J_{\hat{f}}^T = J_{\hat{f}}^T(c^{q+1}) - J_{\hat{f}}^T(c^q)$
- Define $(H_{\hat{f}}^{-T})^{q+1} = \left(I - \frac{\Delta c^q \Delta J_{\hat{f}}}{(\Delta c^q)^T \Delta J_{\hat{f}}^T} \right) (H_{\hat{f}}^{-T})^q \left(I - \frac{\Delta J_{\hat{f}}^T (\Delta c^q)^T}{(\Delta c^q)^T \Delta J_{\hat{f}}^T} \right) + \frac{\Delta c^q (\Delta c^q)^T}{(\Delta c^q)^T \Delta J_{\hat{f}}^T}$
- So that $(H_{\hat{f}}^{-T})^{q+1} \Delta J_{\hat{f}}^T = \Delta c^q$
- Then, solving $H_{\hat{f}}^T(c^{q+1}) \Delta c^{q+1} = -J_{\hat{f}}^T(c^{q+1})$ is replaced with defining the search direction by $\Delta c^{q+1} = -(H_{\hat{f}}^{-T})^{q+1} J_{\hat{f}}^T(c^{q+1})$

L-BFGS (Limited Memory BFGS)

- BFGS stores a dense size $n \times n$ approximation to the inverse Hessian
 - This can become unwieldy for large problems
 - Smarter storage can be accomplished leveraging vectors that describe the outer products; however, the number of vectors still grows with q
- L-BFGS instead estimates the inverse Hessian using only a few vectors
 - often less than 10 vectors (**vectors, vector spaces, not matrices**)
- This makes it quite popular for machine learning
- On optimization methods for deep learning, Andrew Ng et al., ICML 2011
 - “we show that more sophisticated off-the-shelf optimization methods such as Limited memory BFGS (L-BFGS) and Conjugate gradient (CG) with line search can significantly simplify and speed up the process of pretraining deep algorithms”

Nonlinear Least Squares (ML relevancy)

- Minimize a cost function of the form: $\hat{f}(c) = \frac{1}{2} \tilde{f}^T(c) \tilde{f}(c)$
- Recall from Unit 13:
 - Determine parameters c that make $f(x, y, c) = 0$ best fit the training data, i.e. that make $\|f(x_i, y_i, c)\|_2 = \sqrt{f(x_i, y_i, c)^T f(x_i, y_i, c)}$ close to zero for all i
 - Combining all (x_i, y_i) , minimize $\sqrt{\sum_i f(x_i, y_i, c)^T f(x_i, y_i, c)}$ or $\sum_i f(x_i, y_i, c)^T f(x_i, y_i, c)$
 - Minimize $\hat{f}(c) = \sum_i f(x_i, y_i, c)^T f(x_i, y_i, c)$
- Let m be the number of data points and r the output size of $f(x, y, c)$
- Define $\tilde{f}(c)$ by stacking the r outputs of $f(x, y, c)$ consecutively m times, so that the vector valued output of $\tilde{f}(c)$ is length $m * r$
- Then, $\hat{f}(c) = \sum_i f(x_i, y_i, c)^T f(x_i, y_i, c) = \tilde{f}^T(c) \tilde{f}(c)$
 - Multiplying by $\frac{1}{2}$ doesn't change the minimum

Nonlinear Least Squares (Critical Points)

- Minimize $\hat{f}(c) = \frac{1}{2} \tilde{f}^T(c) \tilde{f}(c)$
- Jacobian matrix of \tilde{f} is $J_{\tilde{f}}(c) = \begin{pmatrix} \frac{\partial \tilde{f}}{\partial c_1}(c) & \frac{\partial \tilde{f}}{\partial c_2}(c) & \dots & \frac{\partial \tilde{f}}{\partial c_n}(c) \end{pmatrix}$

- Critical points have $J_{\hat{f}}^T(c) = \begin{pmatrix} \tilde{f}^T(c) \frac{\partial \tilde{f}}{\partial c_1}(c) \\ \tilde{f}^T(c) \frac{\partial \tilde{f}}{\partial c_2}(c) \\ \vdots \\ \tilde{f}^T(c) \frac{\partial \tilde{f}}{\partial c_n}(c) \end{pmatrix} = J_{\tilde{f}}^T(c) \tilde{f}(c) = 0$

Gauss Newton

- $J_{\tilde{f}}^T(c)\tilde{f}(c) = 0$ becomes $J_{\tilde{f}}^T(c)(\tilde{f}(c^q) + J_{\tilde{f}}(c^q)\Delta c^q + \dots) = 0$
 - Using the Taylor series: $\tilde{f}(c) = \tilde{f}(c^q) + J_{\tilde{f}}(c^q)\Delta c^q + \dots$
- Eliminating high order terms: $J_{\tilde{f}}^T(c)(\tilde{f}(c^q) + J_{\tilde{f}}(c^q)\Delta c^q) \approx 0$
- Evaluating $J_{\tilde{f}}^T$ at c^q gives $J_{\tilde{f}}^T(c^q)J_{\tilde{f}}(c^q)\Delta c^q \approx -J_{\tilde{f}}^T(c^q)\tilde{f}(c^q)$
- Compare to $H_{\hat{f}}^T(c^q)\Delta c^q = -J_{\hat{f}}^T(c^q)$ and note that $J_{\hat{f}}^T(c) = J_{\tilde{f}}^T(c)\tilde{f}(c)$
- This implies an estimate $H_{\hat{f}}^T(c^q) \approx J_{\tilde{f}}^T(c^q)J_{\tilde{f}}(c^q)$

Gauss Newton (QR approach)

- The Gauss Newton equations $J_{\tilde{f}}^T(c^q)J_{\tilde{f}}(c^q)\Delta c^q = -J_{\tilde{f}}^T(c^q)\tilde{f}(c^q)$ are the normal equations for $J_{\tilde{f}}(c^q)\Delta c^q = -\tilde{f}(c^q)$
- Thus, (instead) solve $J_{\tilde{f}}(c^q)\Delta c^q = -\tilde{f}(c^q)$ via any least squares (QR) and minimum norm approach
- Note that setting the second factor in $J_{\tilde{f}}^T(c)(\tilde{f}(c^q) + J_{\tilde{f}}(c^q)\Delta c^q) \approx 0$ to zero leads to $\tilde{f}(c^q) + J_{\tilde{f}}(c^q)\Delta c^q = 0$, which is the same $J_{\tilde{f}}(c^q)\Delta c^q = -\tilde{f}(c^q)$
- That is, this is a root finding (or simply downhill) approach to the nonlinear system $\tilde{f}(c) = 0$ (aiming to minimize $\hat{f}(c) = \frac{1}{2}\tilde{f}^T(c)\tilde{f}(c)$)

Weighted Gauss Newton

- Recall: Row scaling changes the importance of the equations
- Recall: Thus, it also changes the (unique) least squares solution for any overdetermined degrees of freedom
- Given a diagonal matrix D indicating the importance of various equations
- $DJ_{\tilde{f}}(c^q)\Delta c^q = -D\tilde{f}(c^q)$
- $J_{\tilde{f}}^T(c^q)D^2J_{\tilde{f}}(c^q)\Delta c^q = -J_{\tilde{f}}^T(c^q)D^2\tilde{f}(c^q)$

Gauss Newton (Regularized)

- When concerned about small singular values in $J_{\tilde{f}}(c^q)\Delta c^q = -\tilde{f}(c^q)$, one can add $\epsilon I = 0$ as extra equations (unit 12)
- This results in $(J_{\tilde{f}}^T(c^q)J_{\tilde{f}}(c^q) + \epsilon^2 I)\Delta c^q = -J_{\tilde{f}}^T(c^q)\tilde{f}(c^q)$
- This is often called Levenberg-Marquardt or Damped (Nonlinear) Least Squares

Gradient/Steepest Descent

- Approximate $H_{\hat{f}}^T$ crudely with the identity matrix
- Then $H_{\hat{f}}^T(c^q)\Delta c^q = -J_{\hat{f}}^T(c^q)$ becomes $I\Delta c^q = -J_{\hat{f}}^T(c^q)$
- That is, the search direction is $\Delta c^q = -J_{\hat{f}}^T(c^q) = -\nabla \hat{f}(c^q)$
- This is the steepest descent direction
- No matrix inversion is necessary
- See unit 19

Coordinate Descent

- Coordinate Descent ignores $H_{\hat{f}}^T(c^q)\Delta c^q = -J_{\hat{f}}^T(c^q)$ completely
- Instead Δc^q is set to the various coordinate directions \hat{e}_k
- No matrix inversion is necessary

Unit 19

Descent Methods

Part II Roadmap

- Part I – Linear Algebra (units 1-12) $Ac = b$
 - Part II – Optimization (units 13-20)
 - (units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima
 - (units 17-18) Computing/Avoiding Derivatives
 - (unit 19) Hack 1.0: “I give up” $H = I$ and J is mostly 0 (descent methods)
 - (unit 20) Hack 2.0: “It’s an ODE!?” (adaptive learning rate and momentum)
-
- ```
graph TD; PartI["Part I – Linear Algebra (units 1-12) $Ac = b$ "] -- "linearize" --> Eq[" $Ac = b$ "]; PartI -- "line search" --> Roots["1D roots/minima"]; PartII["Part II – Optimization (units 13-20)"] --> Theory["Theory"]; PartII --> Methods["Methods"];
```

# Recall: Gradient (Unit 9)

- Consider the scalar (output) function  $f(c)$  with multi-dimensional input  $c$
- The Jacobian of  $f(c)$  is  $J(c) = \begin{pmatrix} \frac{\partial f}{\partial c_1}(c) & \frac{\partial f}{\partial c_2}(c) & \cdots & \frac{\partial f}{\partial c_n}(c) \end{pmatrix}$
- The gradient of  $f(c)$  is  $\nabla f(c) = J^T(c) = \begin{pmatrix} \frac{\partial f}{\partial c_1}(c) \\ \frac{\partial f}{\partial c_2}(c) \\ \vdots \\ \frac{\partial f}{\partial c_n}(c) \end{pmatrix}$
- In 1D, this is the usual  $f'(c)$

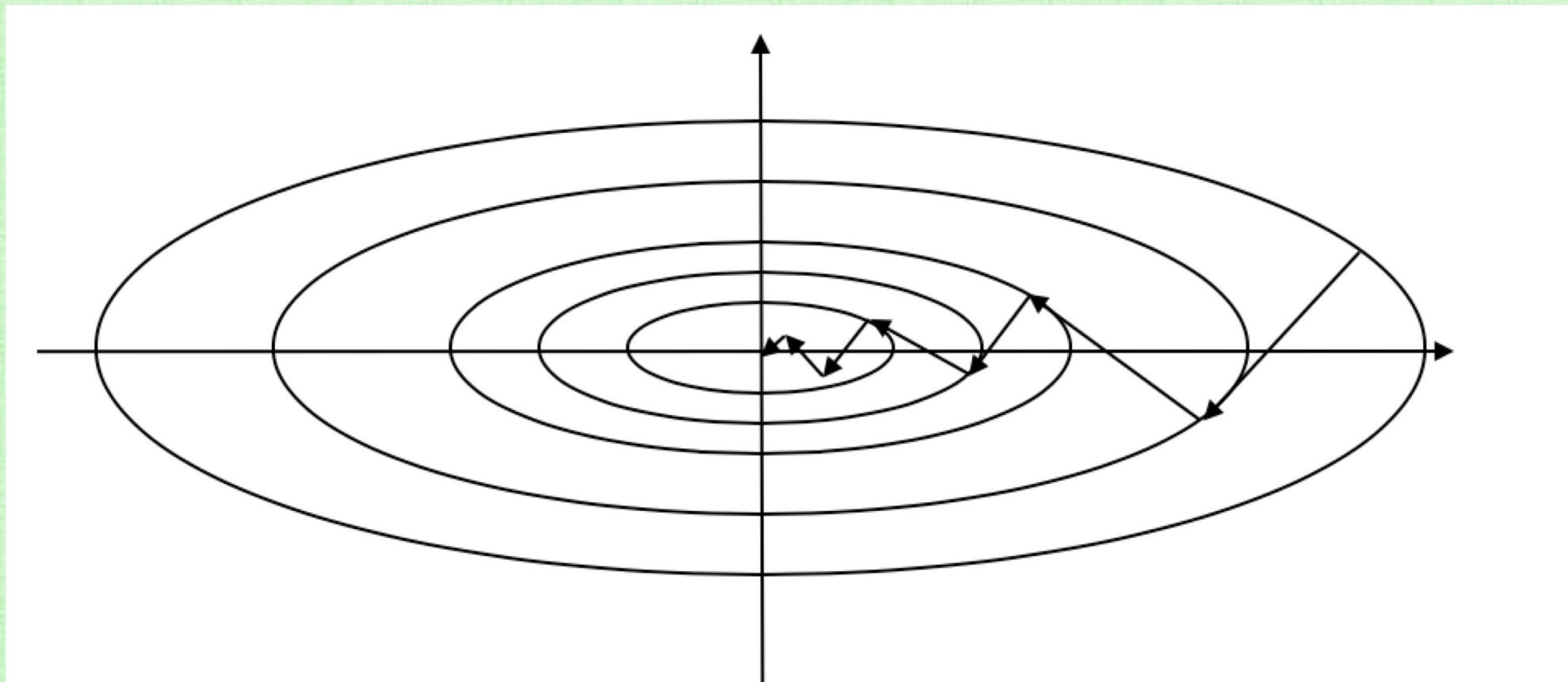
# Gradient/Steepest Descent

- Given a cost function  $\hat{f}(c)$ 
  - $\nabla \hat{f}(c)$  is the direction in which  $\hat{f}(c)$  increases the fastest
  - $-\nabla \hat{f}(c)$  is the direction in which  $\hat{f}(c)$  decreases the fastest
- Thus,  $-\nabla \hat{f}(c)$  is considered the direction of steepest descent
- Using  $-\nabla \hat{f}(c)$  as the search direction is known as steepest descent
  - This can be thought of as always “walking in the steepest downhill direction”
  - However, never going uphill can lead to local minima
- Methods that use  $-\nabla \hat{f}(c)$  in various ways are known as gradient descent methods
- Recall (Unit 18) approximating  $H_{\hat{f}}^T \approx I$  in  $H_{\hat{f}}^T(c^q)\Delta c^q = -J_{\hat{f}}^T(c^q)$  lead to steepest descent  $\Delta c^q = -J_{\hat{f}}^T(c^q) = -\nabla \hat{f}(c^q)$

# Steepest Descent for Quadratic Forms

- Recall (Unit 9):
  - Quadratic Form of SPD  $\tilde{A}$  is  $f(c) = \frac{1}{2}c^T \tilde{A}c - \tilde{b}^T c + \tilde{c}$
  - Minimize  $f(c)$  by finding critical points where  $\nabla f(c) = \tilde{A}c - \tilde{b} = 0$
  - That is, solve  $\tilde{A}c = \tilde{b}$  to find the critical point
- Recall (Unit 5):
  - Steepest descent search direction:  $-\nabla f(c) = \tilde{b} - \tilde{A}c = r$
  - $r^q = b - Ac^q$ ,  $\alpha^q = \frac{r^q \cdot r^q}{r^q \cdot Ar^q}$ ,  $c^{q+1} = c^q + \alpha^q r^q$  is iterated until  $r^q$  is small enough
  - The main drawback to steepest descent is that it repeatedly searches in the same directions too often, especially for higher condition number matrices
  - Thus, it takes a long time to converge (so we advocated Conjugate Gradients)

# Steepest Descent for Quadratic Forms



CG (instead) would solve this in 2 steps

# Recall: Nonlinear Least Squares (Unit 18)

- Minimize a cost function of the form:  $\hat{f}(c) = \frac{1}{2} \tilde{f}^T(c) \tilde{f}(c)$
- Recall from Unit 13:
  - Determine parameters  $c$  that make  $f(x, y, c) = 0$  best fit the training data, i.e. that make  $\|f(x_i, y_i, c)\|_2 = \sqrt{f(x_i, y_i, c)^T f(x_i, y_i, c)}$  close to zero for all  $i$
  - Combining all  $(x_i, y_i)$ , minimize  $\sqrt{\sum_i f(x_i, y_i, c)^T f(x_i, y_i, c)}$  or  $\sum_i f(x_i, y_i, c)^T f(x_i, y_i, c)$
  - Minimize  $\hat{f}(c) = \sum_i f(x_i, y_i, c)^T f(x_i, y_i, c)$
- Let  $m$  be the number of data points and  $r$  the output size of  $f(x, y, c)$
- Define  $\tilde{f}(c)$  by stacking the  $r$  outputs of  $f(x, y, c)$  consecutively  $m$  times, so that the vector valued output of  $\tilde{f}(c)$  is length  $m * r$
- Then,  $\hat{f}(c) = \sum_i f(x_i, y_i, c)^T f(x_i, y_i, c) = \tilde{f}^T(c) \tilde{f}(c)$ 
  - Multiplying by  $\frac{1}{2}$  doesn't change the minimum

# Recall: Nonlinear Least Squares (Unit 18)

- Minimize  $\hat{f}(c) = \frac{1}{2} \tilde{f}^T(c) \tilde{f}(c)$
- Jacobian matrix of  $\tilde{f}$  is  $J_{\tilde{f}}(c) = \begin{pmatrix} \frac{\partial \tilde{f}}{\partial c_1}(c) & \frac{\partial \tilde{f}}{\partial c_2}(c) & \dots & \frac{\partial \tilde{f}}{\partial c_n}(c) \end{pmatrix}$

- Critical points have  $J_{\hat{f}}^T(c) = \begin{pmatrix} \tilde{f}^T(c) \frac{\partial \tilde{f}}{\partial c_1}(c) \\ \tilde{f}^T(c) \frac{\partial \tilde{f}}{\partial c_2}(c) \\ \vdots \\ \tilde{f}^T(c) \frac{\partial \tilde{f}}{\partial c_n}(c) \end{pmatrix} = J_{\tilde{f}}^T(c) \tilde{f}(c) = 0$

# Steepest Descent for Nonlinear Least Squares

- Search direction  $-\nabla \hat{f}(c) = -J_{\hat{f}}^T(c) = -J_{\tilde{f}}^T(c)\tilde{f}(c) = \begin{pmatrix} -\tilde{f}^T(c) \frac{\partial \tilde{f}}{\partial c_1}(c) \\ -\tilde{f}^T(c) \frac{\partial \tilde{f}}{\partial c_2}(c) \\ \vdots \\ -\tilde{f}^T(c) \frac{\partial \tilde{f}}{\partial c_n}(c) \end{pmatrix}$
- Recall that  $\tilde{f}(c)$  is the  $r$  outputs of  $f(x_i, y_i, c)$  stacked consecutively  $m$  times, once for each data point
- Thus, each of the  $n$  terms of the form  $-\tilde{f}^T(c) \frac{\partial \tilde{f}}{\partial c_k}(c)$  is a (potentially expensive) sum through  $m * r$  terms

# Gradient Descent for Nonlinear Least Squares

- When there is a lot of data,  $m$  can be extremely large
  - This is exacerbated when the  $\frac{\partial \tilde{f}}{\partial c_k}$  are expensive to compute
- Using all the data is called Batch Gradient Descent
- When only a small subset of the data is used to compute the search direction (ignoring the rest of the data), this is called Mini-Batch Gradient Descent
- When only a single data point is used to compute the search direction (chosen randomly/sequentially), this is called Stochastic Gradient Descent (SGD)

# Unit 20

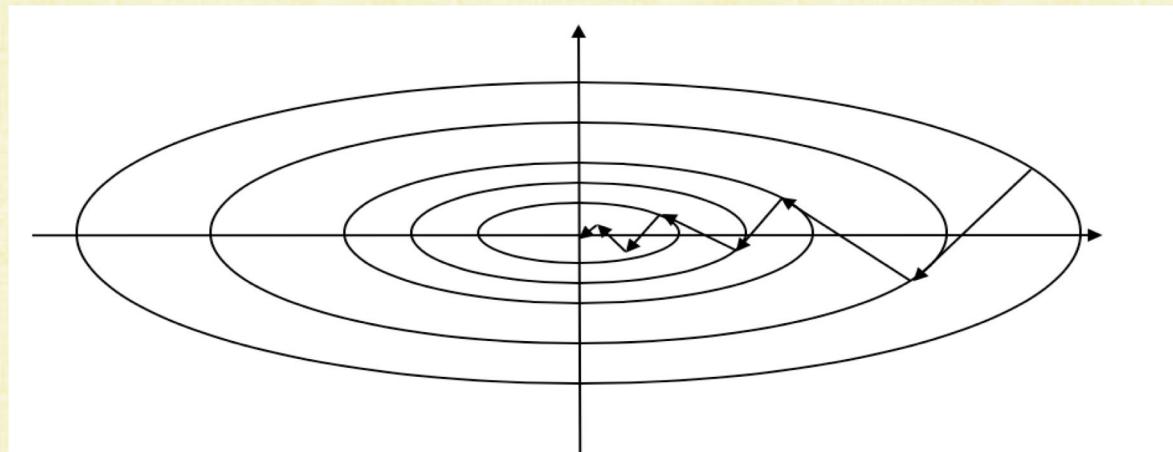
# Momentum Methods

# Part II Roadmap

- Part I – Linear Algebra (units 1-12)  $Ac = b$ 
    - Part II – Optimization (units 13-20)
      - (units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima
      - (units 17-18) Computing/Avoiding Derivatives
      - (unit 19) Hack 1.0: “I give up”  $H = I$  and  $J$  is mostly 0 (descent methods)
      - (unit 20) Hack 2.0: “It’s an ODE!?” (adaptive learning rate and momentum)
- 
- ```
graph TD; PartI["Part I – Linear Algebra (units 1-12)  $Ac = b$ "]; PartII["Part II – Optimization (units 13-20)"]; subgraph Opt ["(units 13-16) Optimization -> Nonlinear Equations -> 1D roots/minima"]; direction TB; subgraph Theory ["Theory"]; direction LR; subgraph Methods ["Methods"]; direction TB; end; subgraph Comp ["(units 17-18) Computing/Avoiding Derivatives"]; end; subgraph Hack19 ["(unit 19) Hack 1.0: ‘I give up’  $H = I$  and  $J$  is mostly 0 (descent methods)"]; end; subgraph Hack20 ["(unit 20) Hack 2.0: ‘It’s an ODE!?’ (adaptive learning rate and momentum)"]; end; PartI -- "linearize" --> Eq[" $Ac = b$ "]; PartI -- "line search" --> LineSearch["line search"]; Eq --> Opt; Theory --> Comp; Theory --> Hack19; Theory --> Hack20; Methods --> Comp; Methods --> Hack19; Methods --> Hack20;
```

Path through Parameter Space

- Optimization solvers, as iterative methods, update the state variable c at each iteration
- For difficult problems (such as neural network training), this is typically done via a 1D line search at each iteration
- The union of all such line searches can be thought of as a path through parameter space



Continuous vs Discrete Path

- Each iteration is a discrete step from one point to another, and connecting them with a 1D line segment is merely a visualization
- Consider the limit as the size of all the line segments goes to zero (and the number of iteration steps equivalently goes to infinity)
- This limit creates the notion of a continuous path in parameter space
- One can parameterize the path/curve with a scalar t (typically called time)
- Then $c(t)$ is a continuous path in parameter space ($c(t)$ is a position)
- Changing the value of t moves the position $c(t)$ along the path
- Differentiating a continuous path gives a time varying velocity: $\frac{dc}{dt}(t)$ or $c'(t)$

Ordinary Differential Equations (ODEs)

- ODEs are equations that describe rates of change
- E.g., $\frac{dc}{dt}(t) = f(t, c(t))$ states that the parameter space velocity is $f(t, c(t))$
- Typically, one “solves” an ODE in order to find the function whose rates of change are described by the ODE
- E.g., given the velocity along the curve $\frac{dc}{dt}(t) = f(t, c(t))$, one aims to determine the curve $c(t)$ itself
- Consider a steepest decent path which always follows the steepest minimizing direction for some cost function $\hat{f}(c)$
 - A suitable velocity is some (positive) scalar multiple of $-\nabla \hat{f}(c(t))$
 - An example ODE describing such a path is $\frac{dc}{dt}(t) = -\nabla \hat{f}(c(t))$

Gradient Flow

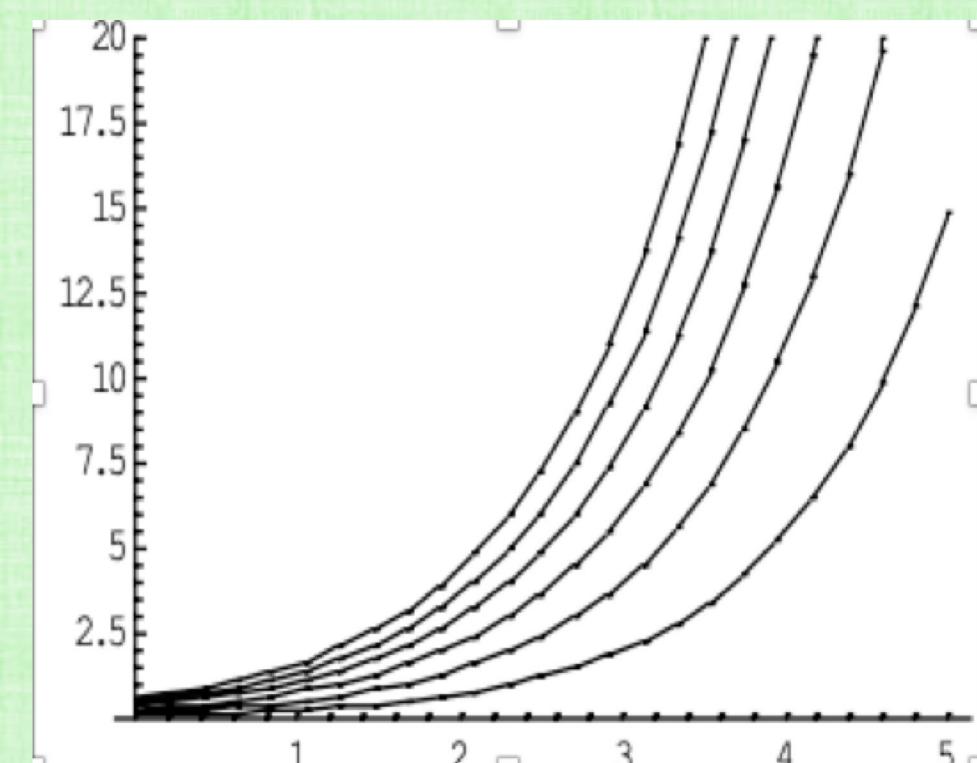
- The ODE for gradient flow, $\frac{dc}{dt}(t) = -\nabla \hat{f}(c(t))$

- Or, in more detail:
$$\begin{pmatrix} \frac{dc_1}{dt}(t) \\ \frac{dc_2}{dt}(t) \\ \vdots \\ \frac{dc_n}{dt}(t) \end{pmatrix} = \begin{pmatrix} -\frac{\partial \hat{f}}{\partial c_1}(c(t)) \\ -\frac{\partial \hat{f}}{\partial c_2}(c(t)) \\ \vdots \\ -\frac{\partial \hat{f}}{\partial c_n}(c(t)) \end{pmatrix}$$

- $c(t)$ is a function of time t that evolves/changes based on the local gradient of the cost function $-\nabla \hat{f}(c(t))$
- This evolution/path follows the direction of steepest descent

Families of Solutions

- ODEs are initial value problems, i.e. their solution depends on the initial (starting) condition



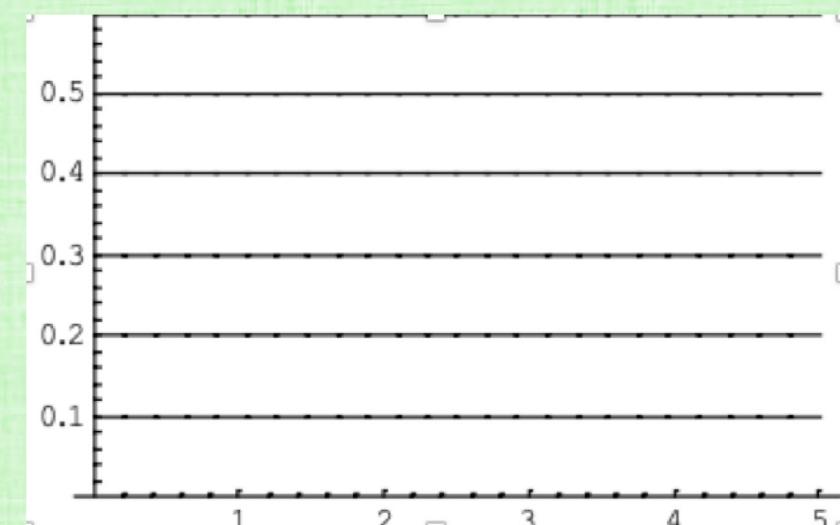
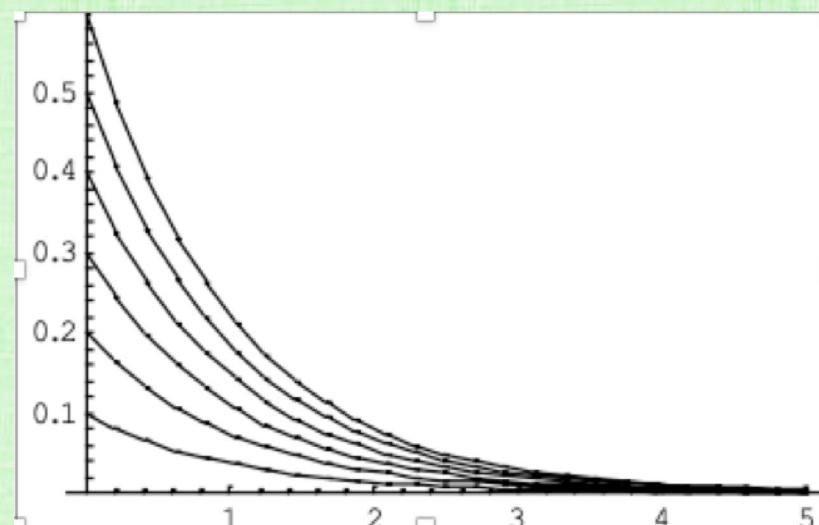
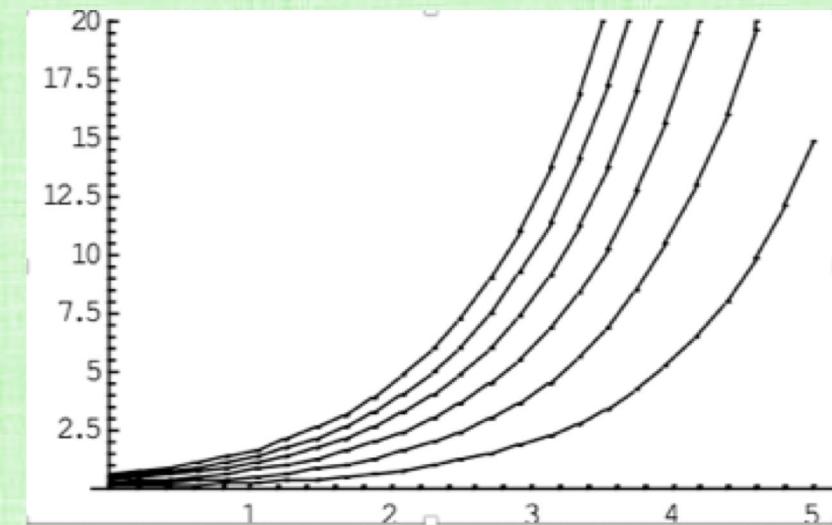
- E.g. $c' = c$ or $\frac{dc}{dt} = c$ or $\frac{dc}{c} = dt$
- $\int_{c_o}^c \frac{1}{c} dc = \int_{t_o}^t dt$ or $\ln c - \ln c_o = t - t_o$
- $\ln \frac{c}{c_o} = t - t_o$ or $\frac{c}{c_o} = e^{t-t_o}$ or $c = c_o e^{t-t_o}$
- Solution $c(t) = c_o e^{t-t_o}$ depends on the initial condition $c(t_o) = c_o$
- Solutions for various values of c_o are shown to the left (where $t_o = 0$)

Solution Families for Gradient Flow

- Following the solution trajectory in gradient flow hopefully leads to a local/global (preferable) minimum of $\hat{f}(c)$
- Various numerical errors will cause perturbations off of the desired trajectory
- Hopefully, perturbed trajectories stay in the same solution family
- Hopefully, solutions in the same family lead to the same local/global minimum
- Sometimes, there are bifurcations of solution trajectories
- In such regions, a perturbation can lead to a family of solutions that end up in a different minimum (which could be local or global, but perhaps not preferred)

Posedness

- Consider $c' = \lambda c$ with solution family $c(t) = c_o e^{\lambda(t-t_o)}$



- $\lambda > 0$, exponential growth, ill-posed
- Small changes in initial conditions (and/or small solver errors) result in large changes to the trajectory
- $\lambda < 0$, exponential decay, well-posed
- Small changes in initial conditions (and/or small solver errors) are damped by converging trajectories
- $\lambda = 0$, constant solution, linearly stable, mildly ill-posed
- Small changes in initial conditions (and/or small solver errors) result in (slow but cumulative) trajectory drift

Posedness

- Consider a system of ODEs $c' = F(t, c)$ with Jacobian matrix $J(t, c) = \frac{\partial F}{\partial c}$
- Since $c(t)$ is time varying, so is $J(t, c(t))$
- Whenever an eigenvalue of $J(t, c(t))$ is positive, the associated part of the solution becomes ill-posed and trajectories begin to (wildly) diverge
 - This can easily pollute the entire solution vector
- Thus, all eigenvalues of $J(t, c(t))$ must be non-positive for all time for the problem to be well-posed
- Moreover, eigenvalues close to zero may be suspect due to numerical errors
- Ill-posedness rapidly leads to solution family bifurcation and (most likely) minima far from what one would expect (and is largely ignored when training networks!)

Stability and Accuracy

- For well-posed ODEs, a numerical approach is considered stable if it does not overflow and produce NaNs (i.e. shoot off to ∞ in parameter space)
- Often, the size of the time step Δt has to be small (enough) to ensure stability
- Otherwise, the method may go unstable
- For well posed ODEs, a stable numerical approach can be analyzed for accuracy to see how well it matches known solutions
- Stability and good accuracy hopefully keep the numerical solution of the ODE within the desired family of solution curves (leading to the preferred minimum)

Forward Euler Method

- Approximate $c' = f(t, c)$ with $\frac{c^{q+1} - c^q}{\Delta t} = f(t^q, c^q)$
- Or recursively as: $c^{q+1} = c^q + \Delta t f(t^q, c^q)$
- Recall: Taylor series $c^{q+1} = c^q + \Delta t f(t^q, c^q) + O(\Delta t^2)$
- So there is an $O(\Delta t^2)$ local truncation error each time step (or iteration)
- Overall, $\frac{t_f - t_0}{\Delta t} = O\left(\frac{1}{\Delta t}\right)$ time steps are taken
- So the total error or global truncation error is $O(\Delta t^2)O\left(\frac{1}{\Delta t}\right) = O(\Delta t)$
- This makes the method 1st order accurate

Runge-Kutta (RK) Methods

- More accurate methods can be constructed similarly (i.e. by consider Taylor series), e.g. RK methods:
- **1st order** $\frac{c^{q+1} - c^q}{\Delta t} = f(t^q, c^q)$ which is forward Euler
- **2nd order** $\frac{c^{q+1} - c^q}{\Delta t} = \frac{1}{2}k_1 + \frac{1}{2}k_2$ where $k_1 = f(t^q, c^q)$ is used in a forward Euler (predictor) update in order to compute $k_2 = f(t^{q+1}, c^q + \Delta t k_1)$
- **4th order** $\frac{c^{q+1} - c^q}{\Delta t} = \frac{1}{6}k_1 + \frac{1}{3}k_2 + \frac{1}{3}k_3 + \frac{1}{6}k_4$ where $k_1 = f(t^q, c^q)$, $k_2 = f\left(t^{q+\frac{1}{2}}, c^q + \frac{\Delta t}{2}k_1\right)$, $k_3 = f\left(t^{q+\frac{1}{2}}, c^q + \frac{\Delta t}{2}k_2\right)$, $k_4 = f(t^{q+1}, c^q + \Delta t k_3)$
 - Again, each term builds on the prior in a predictor style fashion

TVD Runge-Kutta Methods

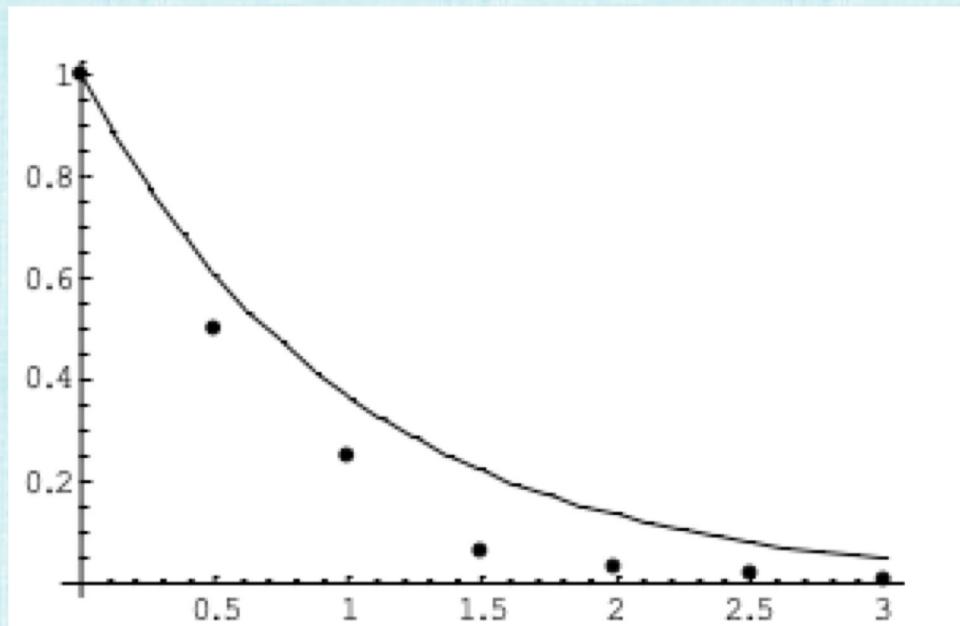
- Combinations of well behaved forward Euler and well behaved averaging
- 1st order same as standard RK and forward Euler
- 2nd order same as standard 2nd order RK, the midpoint rule, the modified Euler method, and Heun's predictor-corrector method
 - Take two forward Euler steps, $\frac{\hat{c}^{q+1} - c^q}{\Delta t} = f(t^q, c^q)$ and $\frac{\hat{c}^{q+2} - \hat{c}^{q+1}}{\Delta t} = f(t^{q+1}, \hat{c}^{q+1})$, and average the initial and final state, $c^{q+1} = \frac{1}{2}c^q + \frac{1}{2}\hat{c}^{q+2}$
- 3rd order Take two Euler steps, but average differently $\hat{c}^{q+\frac{1}{2}} = \frac{3}{4}c^q + \frac{1}{4}\hat{c}^{q+2}$
 - Then take another forward Euler step, $\frac{\hat{c}^{q+\frac{3}{2}} - \hat{c}^{q+\frac{1}{2}}}{\Delta t} = f(t^{q+\frac{1}{2}}, \hat{c}^{q+\frac{1}{2}})$, and average again, $c^{q+1} = \frac{1}{3}c^q + \frac{2}{3}\hat{c}^{q+\frac{3}{2}}$

Stability Analysis

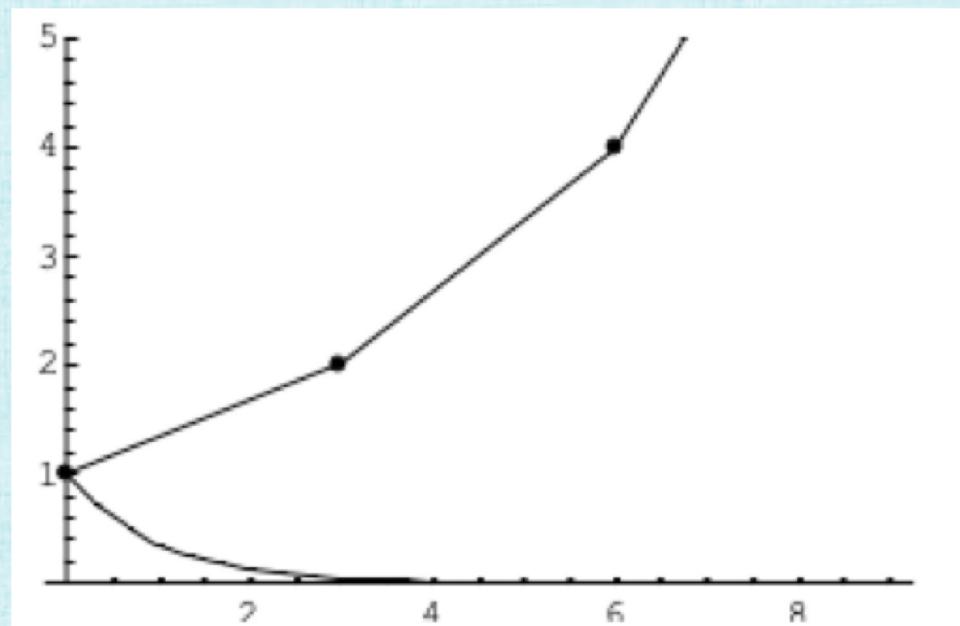
- Consider the model (test) equation $c' = \lambda c$ with a well-posed $\lambda < 0$
- This models how an eigenvalue of a Jacobian matrix might behave
- Forward Euler gives $c^{q+1} = c^q + \Delta t \lambda c^q = (1 + \Delta t \lambda) c^q = (1 + \Delta t \lambda)^{q+1} c^0$
- So the error shrinks and the solutions decays (as expected for $\lambda < 0$) as long as $|1 + \Delta t \lambda| < 1$
- This leads to $-1 < 1 + \Delta t \lambda < 1$ or $-2 < \Delta t \lambda < 0$ or $-\frac{2}{\lambda} > \Delta t > 0$
- Since $\lambda < 0$ and $\Delta t > 0$, one needs $\Delta t < \frac{2}{-\lambda}$ for stability
- This is called a time step restriction

Stability (Example)

- Consider $c' = -c$ with $c(0) = 1$, where $\lambda = -1$ implies $\Delta t < 2$ for stability



- $\Delta t = .5$ is stable
- Iterates (dots) track the solution (curve) well



- $\Delta t = 3$ is unstable
- Iterates (dots) track an incorrect exponentially growing solution
- The actual solution is shown too (decaying)

Gradient Flow

- Using forward Euler on the gradient flow ODE gives: $c^{q+1} = c^q - \Delta t \nabla \hat{f}(c^q)$
- This is the exact same formula utilized for 1D line search $c^{q+1} = c^q + \Delta t \Delta c^q$ in the steepest descent direction, i.e. where $\Delta c^q = -\nabla \hat{f}(c^q)$
- When line search is used, a 1D root/minimization problem is considered to determine the next iterate
- This forward Euler interpretation suggests that one may choose Δt according to various ODE (or other similar) considerations instead

Adaptive Time Stepping

- ODEs utilize either a fixed size Δt or time varying Δt^q , and the latter case is referred to as adaptive time stepping
- The ML community refer to Δt as the learning rate and the time steps as epochs
- Sometimes sub-iterations use only partially valid approximations of $\nabla \hat{f}(c^q)$, e.g. when using mini-batches or SGD for nonlinear least squares (unit 19)
- In those cases, epoch refers to one pass through the entire set of training data, i.e. an epoch allows $\nabla \hat{f}(c^q)$ estimates to see all the data

Adaptive Learning Rates

- Adagrad maintains a separate adaptive learning rate for each parameter, and modifies them based on past gradients computed for that parameter
- Moving more/less in certain directions because of the learning rate (as opposed to the gradient) has the same effect as changing the gradient
- Since this learning rate is based on time history, the optimization method is less localized and presumably/hopefully more robust (better behaved)
- Unfortunately, the learning rates monotonically decrease and often go to zero too quickly stalling out the algorithm
- Adadelta and RMSprop decrease the effect of prior gradients so that the learning rate is not monotonically driven to zero

Implicit Methods

- Used to overcome instabilities seen when using forward Euler (or RK methods) with time steps that are too large
- Implicit methods have either no time step restriction or a very generous one
- However, one typically requires a nonlinear solver to advance each time step
- Sometimes, the nonlinear solver requires more computation than all the small (and simple) time steps forward Euler would require combined together
- Although implicit methods are stable, the large time steps often lead to overly damped solutions or unwanted oscillations

Backward (Implicit) Euler

- $\frac{c^{q+1} - c^q}{\Delta t} = f(t^{q+1}, c^{q+1})$ is 1st order accurate with $O(\Delta t)$ error
- Stability: $\frac{c^{q+1} - c^q}{\Delta t} = \lambda c^{q+1}$ implies $c^{q+1} = \frac{1}{1-\Delta t \lambda} c^q$ where $0 < \frac{1}{1-\Delta t \lambda} < 1$
 - Thus, unconditionally stable, i.e. works for any Δt
- Generally need to solve a nonlinear equation to find c^{q+1} (can be expensive)
- As $\Delta t \rightarrow \infty$, method asymptotes $f(t^{q+1}, c^{q+1}) = 0$ which is ideally a correct steady state solution
 - But, overly damping makes one get there too fast, which is especially undesirable when the high frequencies matter
- Great for stiff problems where high frequencies don't actually contribute much to the solution and thus overly damping them is fine

Implicit Stochastic Gradient Descent (ISGD)

- Used in nonlinear least squares problems to overcome instabilities seen when using forward Euler with time steps that are too large
- Recall: forward Euler is $c^{q+1} = c^q - \Delta t \nabla \hat{f}(c^q)$
- So, backward (implicit) Euler is $c^{q+1} = c^q - \Delta t \nabla \hat{f}(c^{q+1})$
- Since SGD only evaluates the gradient for one piece of data at a time, evaluating the gradient implicitly is a bit less unwieldy

Trapezoidal Rule

- $\frac{c^{q+1} - c^q}{\Delta t} = \frac{f(t^q, c^q) + f(t^{q+1}, c^{q+1})}{2}$ is 2nd order accurate with $O(\Delta t^2)$ error
 - A mix (or average) of forward Euler and backward Euler
- Stability: $\frac{c^{q+1} - c^q}{\Delta t} = \frac{\lambda c^q + \lambda c^{q+1}}{2}$ implies $c^{q+1} = \frac{1 + \frac{\Delta t \lambda}{2}}{1 - \frac{\Delta t \lambda}{2}} c^q$ where $0 < \frac{1 + \frac{\Delta t \lambda}{2}}{1 - \frac{\Delta t \lambda}{2}} < 1$
 - Thus, unconditionally stable, i.e. works for any Δt
- Generally need to solve a nonlinear equation to find c^{q+1} (can be expensive)
- As $\Delta t \rightarrow \infty$, equations become $f(t^{q+1}, c^{q+1}) = -f(t^q, c^q)$ which can cause unwanted oscillations
 - e.g. when $c' = \lambda c$, this is $c^{q+1} = -c^q$ causing oscillations
 - More generally for $c' = f(t, c)$, this is $(c')^{q+1} = -(c')^q$ estimating the derivative as changing sign every iteration (causing oscillations)

Momentum

- Optimization methods can struggle when they are too local (and don't look more holistically at the energy landscape)
- Adaptive learning rates based on time history address this (as seen above)
- Momentum methods also aim to address this
- Momentum methods derive their motivation/name from Newton's Second Law
- Physical objects don't just react to the current state, but carry a time history of their interactions via their momentum
- In particular, the current forces applied to an object are combined with all previous forces to obtain the current trajectory/velocity

Newton's Second Law

- Kinematics describe position $X(t)$ and velocity $V(t)$ as function of time t via $\frac{dX}{dt}(t) = V(t)$ or $X'(t) = V(t)$
 - Gradient flow $\frac{dc}{dt}(t) = -\nabla \hat{f}(c(t))$ is a kinematic equation
- Dynamics describe responses to external stimuli
 - Newton's second law $F(t) = MA(t)$ is a dynamics equation
 - $V'(t) = A(t)$ implies $V'(t) = \frac{F(t)}{M}$ as well as $\frac{d^2X}{dt^2}(t) = X''(t) = \frac{F(t)}{M}$
- Combining kinematics and dynamics gives: $\begin{pmatrix} X'(t) \\ V'(t) \end{pmatrix} = \begin{pmatrix} V(t) \\ \frac{F(t, X(t), V(t))}{M} \end{pmatrix}$
 - Note the potential dependency of forces on position and velocity

Aside: First Order Systems

- Higher order ODEs are often/typically reduced to first order systems
 - E.g. consider: $c'''' = f(t, c, c', c'', c''')$
 - Define new variables: $c_1 = c, c_2 = c', c_3 = c'',$ and $c_4 = c'''$
 - Then $\begin{pmatrix} c_1' \\ c_2' \\ c_3' \\ c_4' \end{pmatrix} = \begin{pmatrix} c_2 \\ c_3 \\ c_4 \\ f(t, c_1, c_2, c_3, c_4) \end{pmatrix}$ is an equivalent first order system
- Similarly, Newton's second law $F = MX''$ was rewritten as $\begin{pmatrix} X' \\ V' \end{pmatrix} = \begin{pmatrix} V \\ F/M \end{pmatrix}$

Momentum Methods

- Rewrite Newton's second law as $\begin{pmatrix} X'(t) \\ MV'(t) \end{pmatrix} = \begin{pmatrix} V(t) \\ F(t, X(t), V(t)) \end{pmatrix}$
 - The second equation augments the momentum with the current forces
 - Then the augmented momentum is used in the first equation (after dividing by mass to get a velocity)
- Interpreting this from an optimization standpoint:
 - Instead of always using the current search direction, one should still be incorporating the effects of prior search directions
- This makes the optimization method less localized, and presumably/hopefully more robust (better behaved)

Gradient Flow

- Split the forward Euler discretization $c^{q+1} = c^q - \Delta t \nabla \hat{f}(c^q)$ into two parts:
$$c^{q+1} = c^q + \Delta t v^q \quad \text{and} \quad v^q = -\nabla \hat{f}(c^q)$$
- Here, v^q is a velocity in parameter space
- Instead of setting the velocity equal to the (negative) gradient, treat gradients as forces that affect the velocity:
$$v^{q+1} = v^q - \Delta t \nabla \hat{f}(c^q)$$

- This results in a forward Euler discretization of $\begin{pmatrix} c'(t) \\ v'(t) \end{pmatrix} = \begin{pmatrix} v(t) \\ -\nabla \hat{f}(c^q) \end{pmatrix}$

“The” ML Momentum Method

- The original momentum method is backward Euler on c and forward Euler on v , i.e. $c^{q+1} = c^q + \Delta t v^{q+1}$ and $v^{q+1} = v^q - \Delta t \nabla \hat{f}(c^q)$
 - Since the second equation can be updated first, the first equation becomes explicit (as opposed to implicit) and doesn't require a special solver
- Combining these into a single equation: $c^{q+1} = c^q + \Delta t v^q - \Delta t^2 \nabla \hat{f}(c^q)$
- Taking liberties to treating Δt and Δt^2 as two separate independent parameters leads to: $c^{q+1} = c^q + \alpha v^q - \beta \nabla \hat{f}(c^q)$
- Setting $\beta = \Delta t$ recovers the original discretization of gradient flow augmented with a new history dependent velocity term: $c^{q+1} = c^q + \alpha v^q - \Delta t \nabla \hat{f}(c^q)$
 - Writing this final equation as $c^{q+1} = c^q + \Delta t v^{q+1}$ illustrates an inconsistent velocity update of $v^{q+1} = \frac{\alpha}{\Delta t} v^q - \nabla \hat{f}(c^q)$

Nesterov Momentum

- Nesterov uses a predictor-corrector approach similar to second order Runge-Kutta
- First, a forward Euler predictor step is taken $\hat{c}^{q+1} = c^q + \Delta t \hat{v}^{q+1}$ using a velocity of $\hat{v}^{q+1} = \frac{\alpha}{\Delta t} v^q$ (instead of $v^{q+1} = \frac{\alpha}{\Delta t} v^q - \nabla \hat{f}(c^q)$ from the last slide)
 - That is, the current gradient information is ignored in the predictor step
 - Simplifying, the predictor step is $\hat{c}^{q+1} = c^q + \alpha v^q$
- Then, the gradient is evaluated at this new location \hat{c}^{q+1} and used in the original momentum equations: $c^{q+1} = c^q + \Delta t v^{q+1}$ and $v^{q+1} = \frac{\alpha}{\Delta t} v^q - \nabla \hat{f}(\hat{c}^{q+1})$
 - As a single equation $c^{q+1} = c^q + \alpha v^q - \Delta t \nabla \hat{f}(\hat{c}^{q+1})$
 - Once again, there is an inconsistent velocity update $v^{q+1} = \frac{\alpha}{\Delta t} v^q - \nabla \hat{f}(\hat{c}^{q+1})$

Physics/ODE Consistency

- As an ODE from physics, classical numerical analysis dictates that the correct solution/path should be obtained as $\Delta t \rightarrow 0$
 - Although $c^{q+1} = c^q + \Delta t v^{q+1}$ properly resolves the path based on a velocity, $v^{q+1} = \frac{\alpha}{\Delta t} v^q - \nabla \hat{f}(\tilde{c})$ (where \tilde{c} is either c^q or \hat{c}^{q+1}) is problematic
- Revert to where we took liberties with $c^{q+1} = c^q + \alpha v^q - \beta \nabla \hat{f}(\tilde{c})$
- This time, choose $\beta = \hat{\beta} \Delta t^2$ (instead of $\beta = \Delta t$) to obtain a velocity equation of the form $v^{q+1} = \frac{\alpha}{\Delta t} v^q - \Delta t \hat{\beta} \nabla \hat{f}(\tilde{c})$
- Setting $\alpha = \Delta t$ leads to a consistent physical system $v^{q+1} = v^q - \Delta t \hat{\beta} \nabla \hat{f}(\tilde{c})$ with scalar $\hat{\beta} > 0$ determining the strength of the steepest descent force
 - Like all physical systems, forces should be independent of Δt , and should accumulate to the same $O(1)$ net effect in $O(1)$ time, regardless of Δt

Adam

- Mix ideas from adaptive learning rates and momentum methods:
 - Adaptive learning rate for each parameter (uses squared gradients to scale the learning rate like RMSprop)
 - Uses a moving average of the gradient like momentum methods
- AdaMax variant uses the L^∞ norm instead of the L^2 norm
- Nadam variant uses Nesterov momentum for the moving averages
- The original Adam paper had impressive results, which were duplicated by others, and the method has been quite popular
- More recent work is finding that Adam might converge quicker than SGD with momentum, but sometimes quicker to a worse solution (i.e. some are going back to SGD)
 - Still a lot to do!

Constant Acceleration Equations

- Taylor expansion: $X^{q+1} = X^q + \Delta t V^q + \frac{\Delta t^2}{2} A^q + O(\Delta t^3)$
- In order to determine X^{q+1} with $O(\Delta t^3)$ accuracy, one only needs V^q with $O(\Delta t^2)$ accuracy and A^q with $O(\Delta t)$ accuracy
- E.g., in the system of equations for Newtons second law, the second equation $V' = F/M$ requires $O(\Delta t)$ less accuracy than the first equation $X' = V$
- The standard kinematic formulas in basic physics use:
 - piecewise constant accelerations A^q
 - piecewise linear velocities $V^{q+1} = V^q + \Delta t A^q$
 - piecewise quadratic positions $X^{q+1} = X^q + \Delta t V^q + \frac{\Delta t^2}{2} A^q$

Newmark Methods

- $X^{q+1} = X^q + \Delta t V^q + \frac{\Delta t^2}{2} ((1 - 2\beta)A^q + 2\beta A^{q+1})$
- $V^{q+1} = V^q + \Delta t ((1 - \gamma)A^q + \gamma A^{q+1})$
- $\beta = \gamma = 0$ constant acceleration equations
- Second order accurate if and only if $\gamma = \frac{1}{2}$, i.e. $V^{q+1} = V^q + \Delta t \frac{A^q + A^{q+1}}{2}$
- $\gamma = \frac{1}{2}, \beta = \frac{1}{4}$ Trapezoidal Rule (on both X and V)
 - $X^{q+1} = X^q + \Delta t V^q + \frac{\Delta t^2}{4} (A^q + A^{q+1})$ becomes $X^{q+1} = X^q + \Delta t V^q + \frac{\Delta t}{2} (V^{q+1} - V^q)$ or
 $X^{q+1} = X^q + \Delta t \frac{V^q + V^{q+1}}{2}$
- $\gamma = \frac{1}{2}, \beta = 0$ Central Differencing ($X^{q+1} = X^q + \Delta t V^q + \frac{\Delta t^2}{2} A^q$)

Central Differentiating

- $X^{q+1} = X^q + \Delta t V^q + \frac{\Delta t^2}{2} A^q$ and $V^{q+1} = V^q + \Delta t \frac{A^q + A^{q+1}}{2}$
- Adding $X^{q+2} = X^{q+1} + \Delta t V^{q+1} + \frac{\Delta t^2}{2} A^{q+1}$ to $X^{q+1} = X^q + \Delta t V^q + \frac{\Delta t^2}{2} A^q$ gives
$$X^{q+2} - X^q = \Delta t(V^q + V^{q+1}) + \frac{\Delta t^2}{2}(A^q + A^{q+1}) = \Delta t(V^q + V^{q+1}) + \Delta t(V^{q+1} - V^q) = 2\Delta t V^{q+1}$$
- So $V^{q+1} = \frac{X^{q+2} - X^q}{2\Delta t}$ (a second order accurate central difference)
- Subtracting (same equations) gives $X^{q+2} - 2X^{q+1} + X^q = \Delta t(V^{q+1} - V^q) + \frac{\Delta t^2}{2}(A^{q+1} - A^q) = \frac{\Delta t^2}{2}(A^q + A^{q+1}) + \frac{\Delta t^2}{2}(A^{q+1} - A^q) = \Delta t^2 A^{q+1}$
- So $A^{q+1} = \frac{X^{q+2} - 2X^{q+1} + X^q}{\Delta t^2}$ (a second order accurate central difference)

Staggered Position and Velocity

- Update position with a staggered velocity $X^{q+1} = X^q + \Delta t V^{q+\frac{1}{2}}$
- Using averaging $V^{q+1} = \frac{V^{q+\frac{1}{2}} + V^{q+\frac{3}{2}}}{2}$ which still equals $\frac{X^{q+2} - X^q}{2\Delta t}$ as desired
- $A^{q+1} = \frac{(X^{q+2} - X^{q+1}) - (X^{q+1} - X^q)}{\Delta t^2} = \frac{V^{q+\frac{3}{2}} - V^{q+\frac{1}{2}}}{\Delta t}$
- This last term is equal to both $\frac{V^{q+1} - V^{q+\frac{1}{2}}}{(\Delta t/2)}$ and $\frac{V^{q+\frac{3}{2}} - V^{q+1}}{(\Delta t/2)}$
- So $V^{q+1} = V^{q+\frac{1}{2}} + \frac{\Delta t}{2} A^{q+1}$ and $V^{q+\frac{3}{2}} = V^{q+1} + \frac{\Delta t}{2} A^{q+1}$
- The second equation shifted one index is $V^{q+\frac{1}{2}} = V^q + \frac{\Delta t}{2} A^q$

Staggered Central Differencing

- $V^{q+\frac{1}{2}} = V^q + \frac{\Delta t}{2} A(X^q, V^q)$ and $X^{q+1} = X^q + \Delta t V^{q+\frac{1}{2}}$ are explicit
- $V^{q+1} = V^{q+\frac{1}{2}} + \frac{\Delta t}{2} A(X^{q+1}, V^{q+1})$ is explicit in X but implicit in V
- Position based forces (e.g. elasticity) are typically nonlinear making them hard to invert (good that we don't have to), whereas velocity based forces (e.g. damping) are typically linear making them easier to invert (which we do)
- Position based forces are often important for material behavior (good we don't overdamp them), whereas velocity based damping doesn't suffer much from increased damping (which we do if we switch from trapezoidal rule to backward Euler in the last step, i.e. $V^{q+1} = V^q + \Delta t A(X^{q+1}, V^{q+1})$)
- Position based forces don't require too stringent a time step restriction (good, because we need one), whereas velocity based forces typically require a very small time step restriction (which we ignore with an implicit solver)

Appendix

Notation

Unit 1: Intro

- x, y, z data inputs/outputs
- $i = 1, m$ subscript enumerates data (and thus rows of A)
- f function of the data
- $\hat{x}, \hat{y}, \hat{z}, \hat{f}, \hat{\varphi}$ inference/approximation of same variables
- c unknown parameters to characterize functions
- $k = 1, n$ subscript enumerates c (and thus columns of A)
- A matrix (I identity), b right hand side (y used when it is the actual data y)
- a_k column of A
- Quadratic Formula slide: uses standard notation for all variables
- ϕ basis functions
- θ pose parameters, φ collection of all vertex positions for a triangle surface mesh
- Cloth Slides: S, D functions, u, v texture space, n normal direction, I image data, h interpolation function

Unit 2: Linear Systems

- a_{ik} elements of A
- A^T transpose, A^{-1} inverse
- \hat{e}_i standard basis vectors
- Gaussian Elimination slides m_{ik} special column, M_{ik} , L_{ik} special matrices
- $I_{m \times m}$ size $m \times m$ identity
- U upper triangular matrix, L lower triangular matrix
- \hat{c} transformed version of c
- P permutation matrix (with its own special notation)

Unit 3: Understanding Matrices

- λ eigenvalue (scalar)
- v eigenvector, u right eigenvector (both column vectors)
- α scalar
- * superscript is complex conjugate (for imaginary numbers)
- $i = \sqrt{-1}$ when dealing with complex numbers
- \hat{c}, \hat{b} perturbed or transformed b, c
- \hat{A}^{-1}, \hat{I} approximate version A^{-1}, I
- U, V orthogonal (for SVD)
- Σ diagonal (not necessarily square, potentially zeros on diagonal)
- σ singular values

Unit 4: Special Matrices

- v, u column vectors
- u_k, v_k columns of U, V
- Λ diagonal matrix of eigenvalues
- l_{ik} element of L
- \hat{A} approximation of A

Unit 5: Iterative Solvers

- q superscript, integer for sequences/iterations (iterative solvers)
- ϵ small number
- t time
- X, V position and velocity
- r, e residual and error (column vectors)
- s search direction
- \bar{S} column vector
- β scalar

Unit 6: Local Approximations

- h scalar (relatively small)
- $f^{(p)}$ parenthesis (integer) indicate taking p derivatives
- f' and f'' one derivative and two derivatives
- Cubic Splines Slide: special notation
- p integer, polynomial degree, order of accuracy, etc.
- w weighting function

Unit 7: Curse of Dimensionality

- A, V area and volume
- r radius
- N integer, number of sample points
- \vec{x} vector of data input to a function

Unit 8: Least Squares

- False Statements (first slide): a, b scalars
- D, \hat{D} diagonal matrices

Unit 9: Basic Optimization

- F system of functions (output is a vector not a scalar)
- ∂ partial derivative symbol
- J Jacobian matrix of all first partial derivatives
- F' Jacobian of F
- ∇f gradient of scalar function f (Jacobian transposed)
- H matrix of all second partial derivatives of scalar function f (Jacobian of gradient transposed)
- c^* critical point (special value of c)
- $\tilde{A}, \tilde{b}, \tilde{c}$ matrix, and two vectors

Unit 10: Solving Least Squares

- $\hat{\Sigma}$ diagonal invertible matrix (no zeros on the diagonal)
- $I_{n \times n}$ stresses the size of the identity as $n \times n$
- \hat{b}_r, \hat{b}_z sub-vectors of \hat{b} of shorter length (range and zero abbreviations)
- Q orthogonal matrix
- q_k column of Q
- R upper triangular matrix
- r_{ik} entry of R
- \tilde{Q} submatrix
- Householder slides: \hat{v} normal vector, H householder matrix, a column vector (all this notation is specialized)

Unit 11: Zero Singular Values

- c_r, c_z sub-vectors of \hat{c} of shorter length (range and zero abbreviations)
- A^+ pseudo-inverse of A
- T matrix (for similarity transforms)
- Power Method Slides: A^q and λ^q are A and λ raised to the q power (not an iteration as is the case for other q 's on these slides)

Unit 12: Regularization

- c^* is an initial guess for c
- r used in its geometric series capacity (a scalar)
- θ angle between two vectors
- C, C^* curves (vertices connected by line segments)

Unit 13: Optimization

- f briefly is allowed to be either vector valued (or stay scalar)
- \hat{f} becomes the (scalar) cost function for optimization
- F system of functions (gradient in the case of optimization)

Unit 14: Nonlinear Systems

- c^* is a point to linearize about
- d is for the standard derivative
- t is an arbitrary variable
- dt is a differential
- Δ finite size difference
- g scalar function (that determines the line search parameter)

Unit 15: Root Finding

- \hat{g} modified g
- t search parameter in 1D, replacing α
- t^* converged solution
- \hat{t} particular t
- C scalar
- p integer (power)
- g' derivative of g
- t_L, t_R interval bounds
- t_M inteval midpoint

Unit 16: 1D Optimization

- t_{min}, t_{M1}, t_{M2} more t values
- s scalar (interval size)
- f, τ scalars between 0 and 1
- H_F is a 3rd order tensor of 2nd derivatives
- $OMG_{\hat{f}}$ is a 3rd order of 3rd derivatives

Unit 17: Computing Derivatives

- TBA

Unit 18: Avoiding Derivatives

- TBA

Unit 19: Descent Methods

- TBA

Unit 20: Momentum Methods

- TBA