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)

linearize

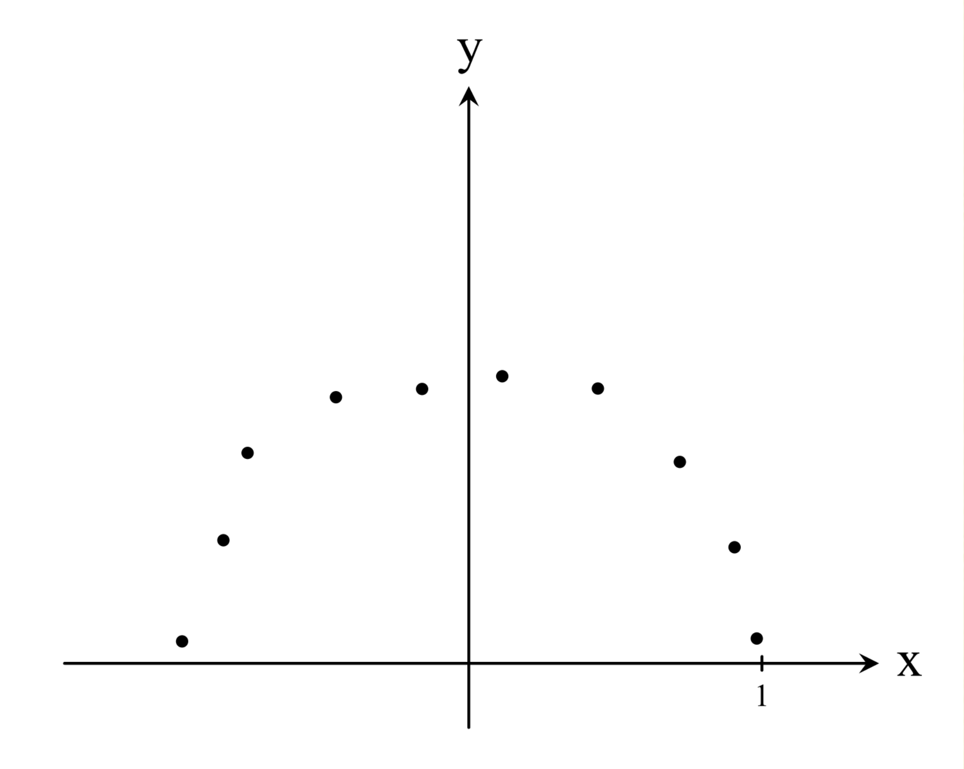
line search

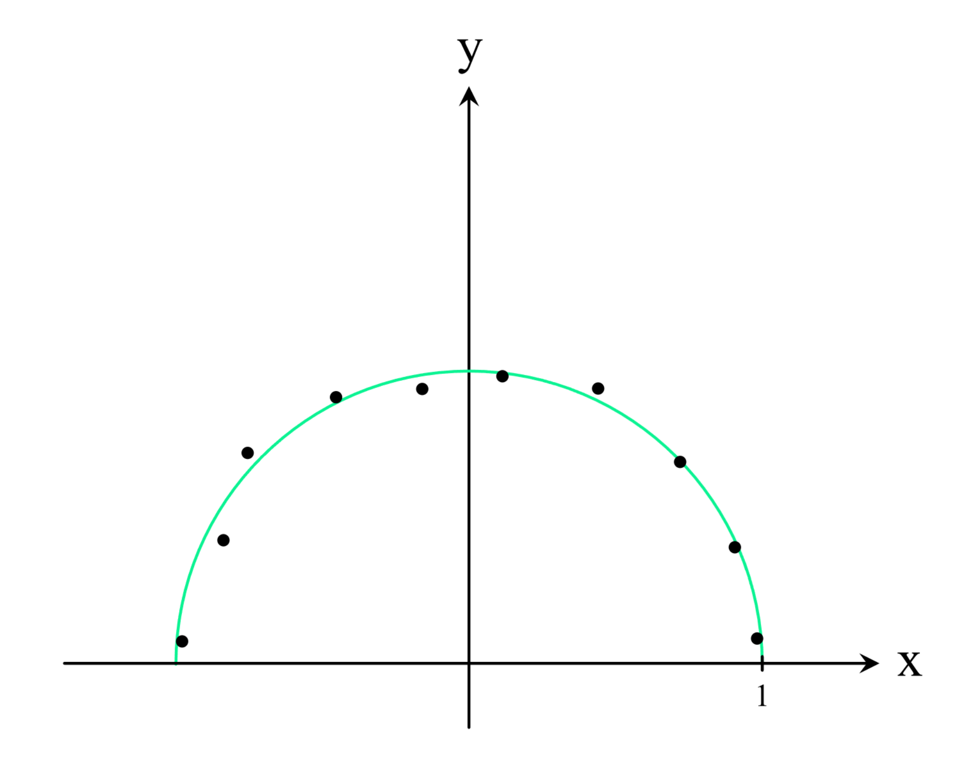
Theory

Methods

Approximating Functions

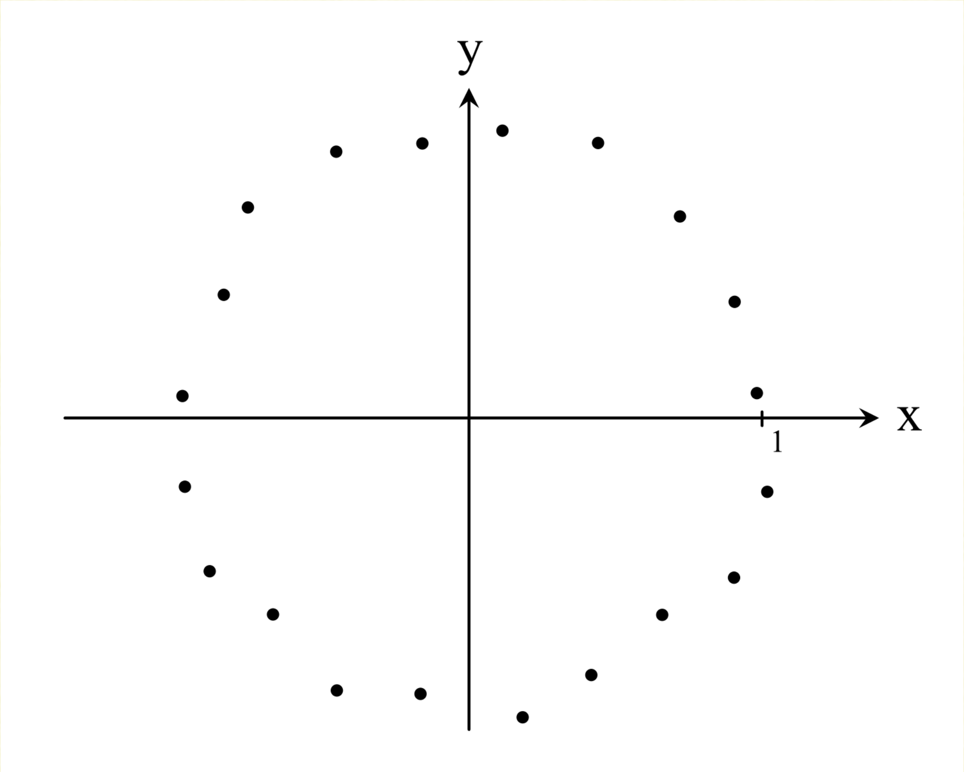
- Consider the (x_i, y_i) data shown below
- Here, $y = \sqrt{1 - x^2}$ looks like a good approximation

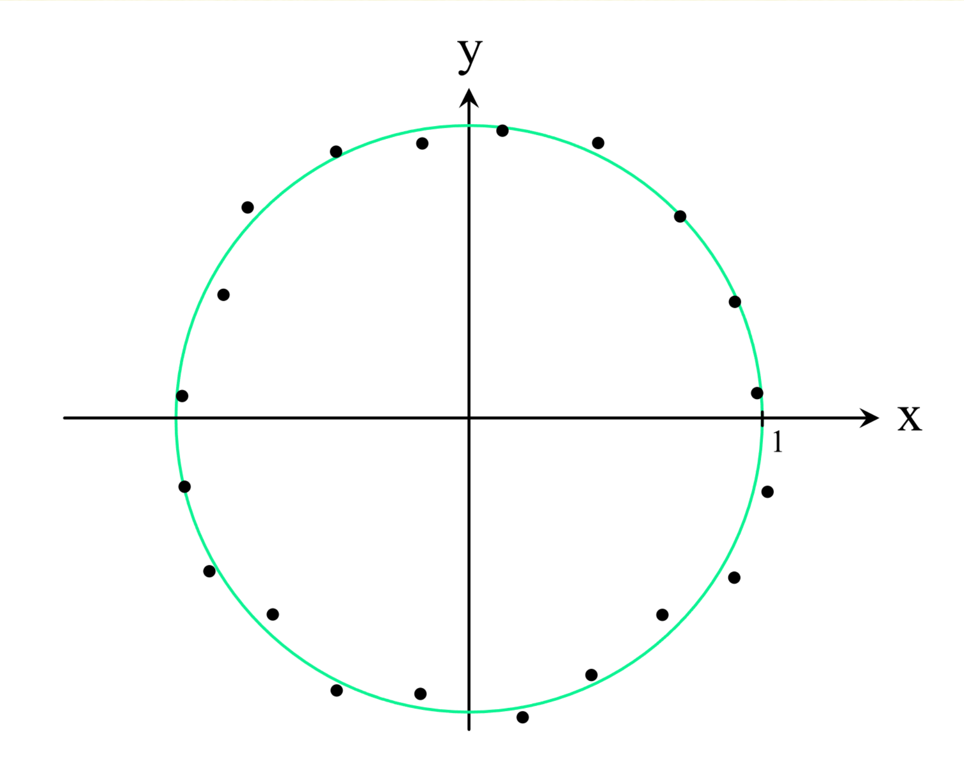




Approximating Functions

- Consider the (x_i, y_i) data shown below
- Here, $x^2 + y^2 = 1$ looks like a good approximation (fails the vertical line test)





Approximating Functions

- A function does not need to be explicit in y
- Any relationship between x and y is fine, 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 (a <u>model</u> for f)
  - E.g., f could be all possible circles $(x - c_1)^2 + (y - c_2)^2 - c_3^2 = 0$ 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 circles, or polynomials, or a <u>network architecture</u>, etc.
- <u>Determine parameters</u> c that make $f(x, y; c) = 0$ <u>best fit the data</u>, 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; for vectors, a norm needs to be chosen for $\|f(x_i,y_i;c)\|$, e.g. L^1, L^2, L^∞, “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 all of these together as well
  - E.g., $\sqrt{\sum_i \|f(x_i,y_i;c)\|_2^2} = \sqrt{\sum_i f(x_i,y_i;c)^T f(x_i,y_i;c)}$
- Minimize $\sqrt{\sum_i f(x_i,y_i;c)^T f(x_i,y_i;c)}$ or equivalently $\sum_i f(x_i,y_i;c)^T f(x_i,y_i;c)$
- Since all the (x_i,y_i) are known, the <u>cost function</u> 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 <u>Nonlinear Least Squares</u>

Optimization

- Minimize the cost function $\hat{f}(c)$
- Since maximizing $\hat{f}(c)$ is equivalent to minimizing $-\hat{f}(c)$, optimization is typically 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
- When constraints are present, denoted constrained (as opposed unconstrained) optimization
  - Constraints can be 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)

Conditioning

- Recall: Minimizing the residual $r = b - Ac$ with an L^2 norm led to the 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
  - Partial derivatives approaching 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 $\hat{f}(c) = 0$ (i.e. for finding the roots of $\hat{f}(c) = 0$)
  - Can only expect half as many significant digits of accuracy
  - If an error tolerance of ϵ would be used for solving $\hat{f}(c) = 0$, then a weaker (larger) $\sqrt{\epsilon}$ error tolerance is more appropriate for minimizing $\hat{f}(c)$

Nonlinear Systems of Equations

- Critical points have $\frac{\partial \hat{f}}{\partial c_k}(c) = 0$ 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, add a penalty term $\hat{g}^T(c)D\hat{g}(c) \geq 0$ to the cost function and proceed via unconstrained optimization
  - I.e., minimize $\hat{f}(c) + \hat{g}^T(c)D\hat{g}(c)$ via unconstrained optimization

- Various other options also exist:
  - E.g. Add Lagrange multipliers η as new variables, and minimize $\hat{f}(c) + \eta^T\hat{g}(c)$

Lagrange Multipliers

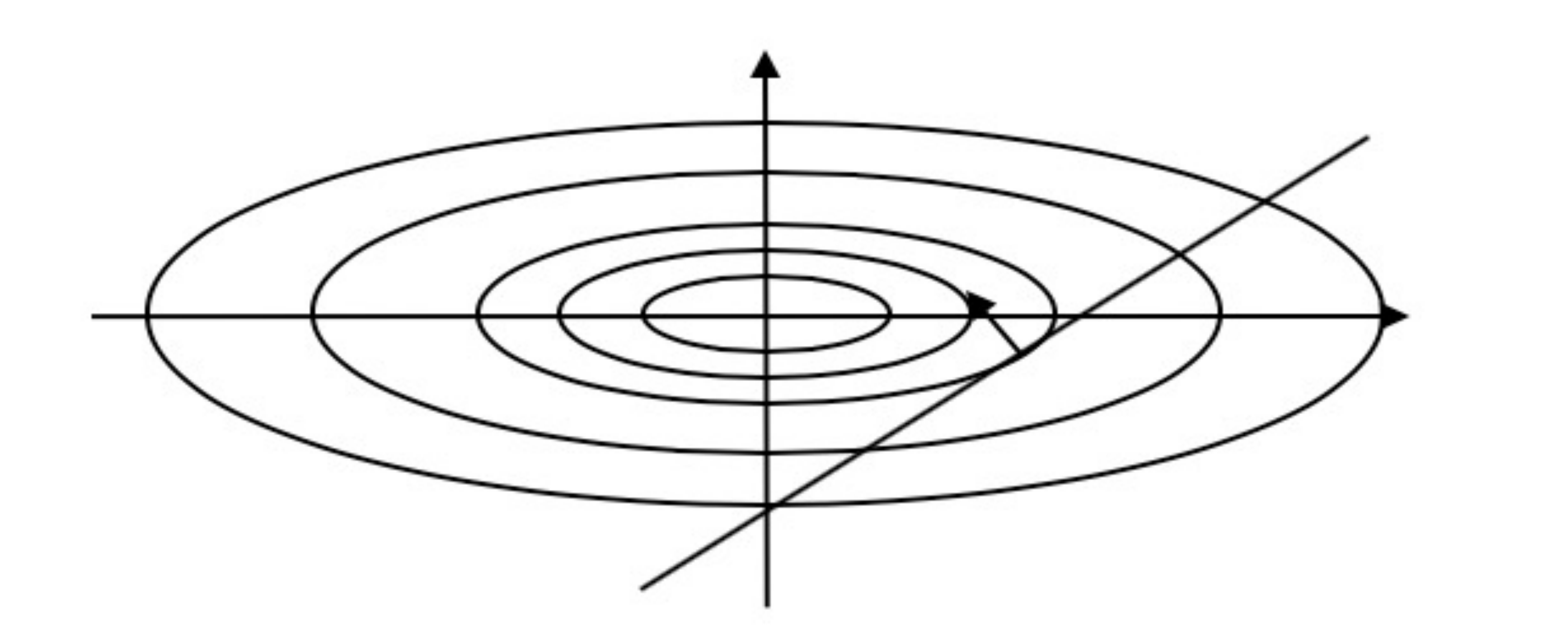
- Minimize $\hat{f}(c) + \eta^T \hat{g}(c)$
- Critical Points: $\nabla\left(\hat{f}(c) + \eta^T \hat{g}(c)\right) = \begin{pmatrix} J_{\hat{f}}^T(c) + J_{\hat{g}}^T(c)\eta \\ \hat{g}(c) \end{pmatrix} = 0$
  - Note how the $\hat{g}(c) = 0$ constraints are automatically satisfied at critical points
- Critical 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 $\hat{f}(c) = \frac{1}{2}c_1^2 + \frac{5}{2}c_2^2$ subject to $\hat{g}(c) = c_1 - c_2 - 1 = 0$
- Or, minimize $\frac{1}{2}c_1^2 + \frac{5}{2}c_2^2 + \eta_1(c_1 - c_2 - 1)$
- Critical Points: $\begin{pmatrix} \begin{pmatrix} c_1 \\ 5c_2 \end{pmatrix} + \begin{pmatrix} 1 \\ -1 \end{pmatrix}\eta_1 \\ c_1 - c_2 - 1 \end{pmatrix} = \begin{pmatrix} c_1 + \eta_1 \\ 5c_2 - \eta_1 \\ c_1 - c_2 - 1 \end{pmatrix} = 0$
- Or, $\begin{pmatrix} 1 & 0 & 1 \\ 0 & 5 & -1 \\ 1 & -1 & 0 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ \eta_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ or $\begin{pmatrix} c_1 \\ c_2 \\ \eta_1 \end{pmatrix} = \begin{pmatrix} 5/6 \\ -1/6 \\ -5/6 \end{pmatrix}$
- The Hessian is $\begin{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix} & \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\ \begin{pmatrix} 1 & -1 \end{pmatrix} & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 5 & -1 \\ 1 & -1 & 0 \end{pmatrix}$

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 $-\nabla\hat{f} = \begin{pmatrix} -5/6 \\ 5/6 \end{pmatrix}$ is perpendicular to the constraint surface (which has $(1,1)$ as the line direction)



Lagrange Multipliers (Example)

- Plug $c_2 = c_1 - 1$ into $\hat{f}(c)$ to get $\frac{1}{2}c_1^2 + \frac{5}{2}(c_1 - 1)^2 = 3c_1^2 - 5c_1 + \frac{5}{2}$, which is a parabola with minimum at $c_1 = \frac{5}{6}$ (as expected)