Linear Systems and Optimization

Jeffrey E. Boyd

UNIVERSITY OF CALGARY

overview

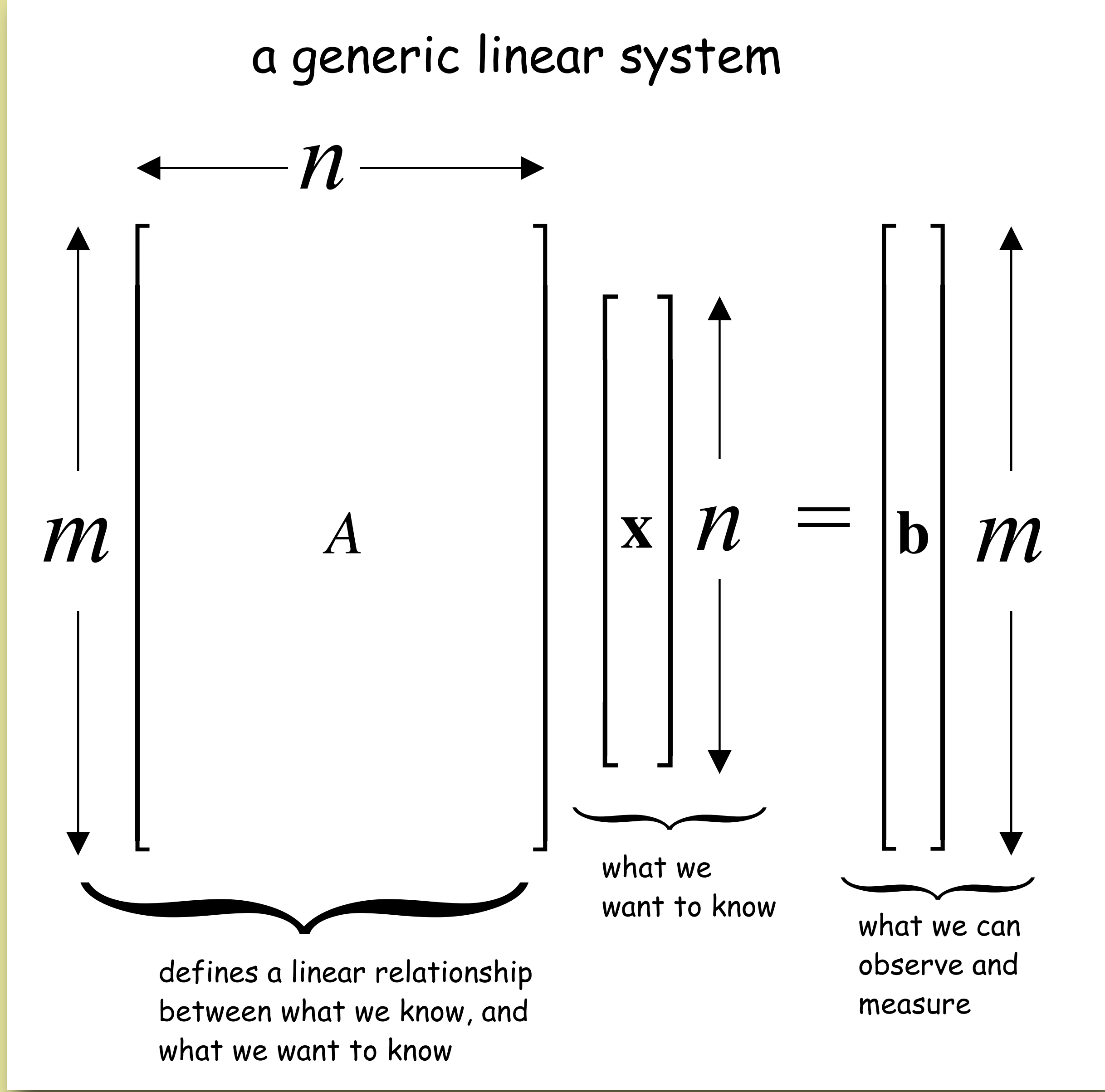
- start with our basic linear system

$$A\mathbf{x}=\mathbf{b}$$

- defines a quadratic function
- for certain forms of A, can solve system by minimizing the quadratic function

generic linear system

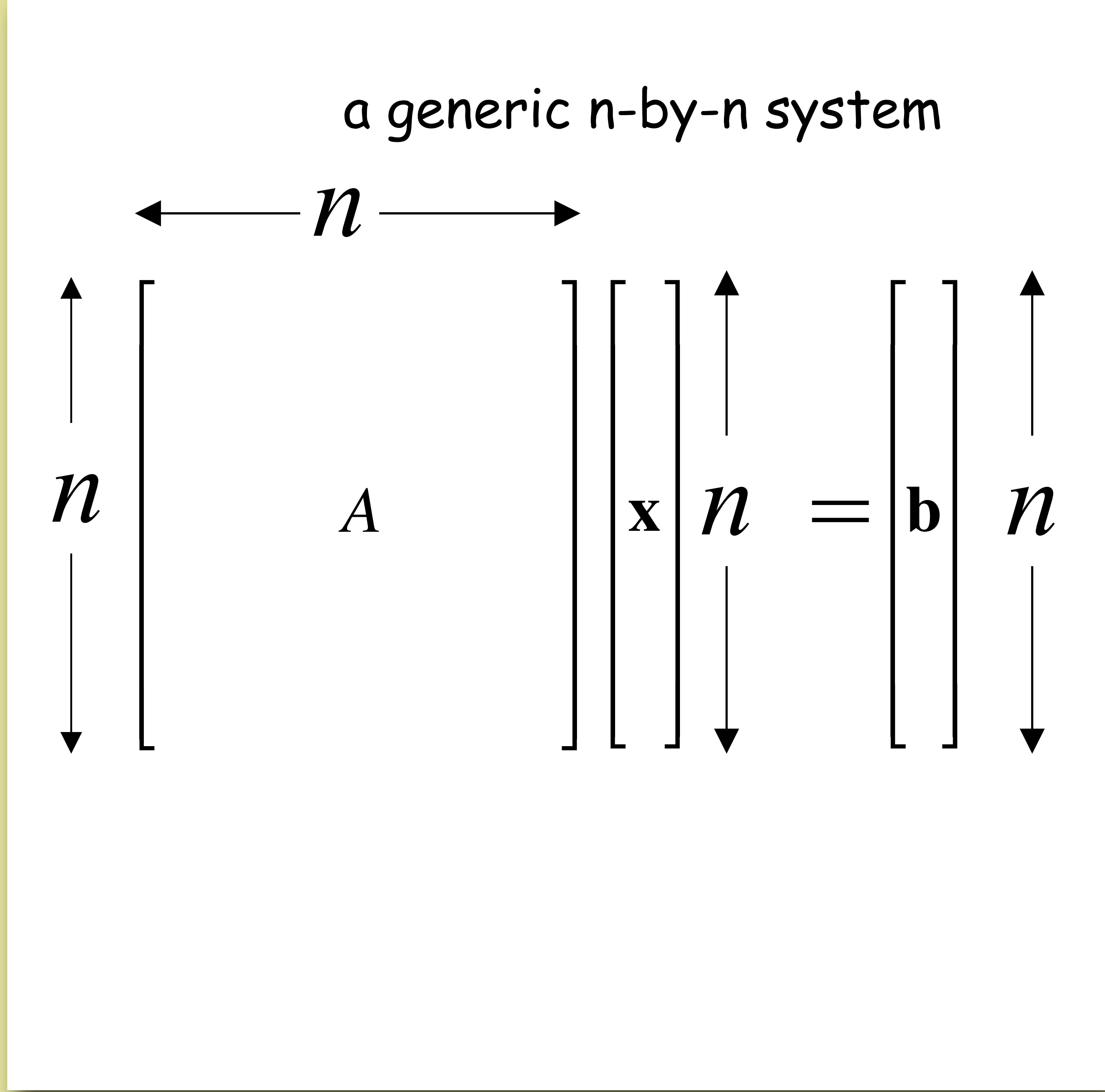
- measure *m* values in **b**.
- want to know the *n* values in **x**.
- *A* defines relationship between **b** and **x**.



UNIVERSITY OF CALGARY

n-by-n system

- if *A* is square we get a generic n-by-n linear system



invert and multiply

- if *A* is square we can invert and multiply to solve
  - *A* must be non-singular

move *A* to other side - invert and multiply

$$A\mathbf{x} = \mathbf{b}$$

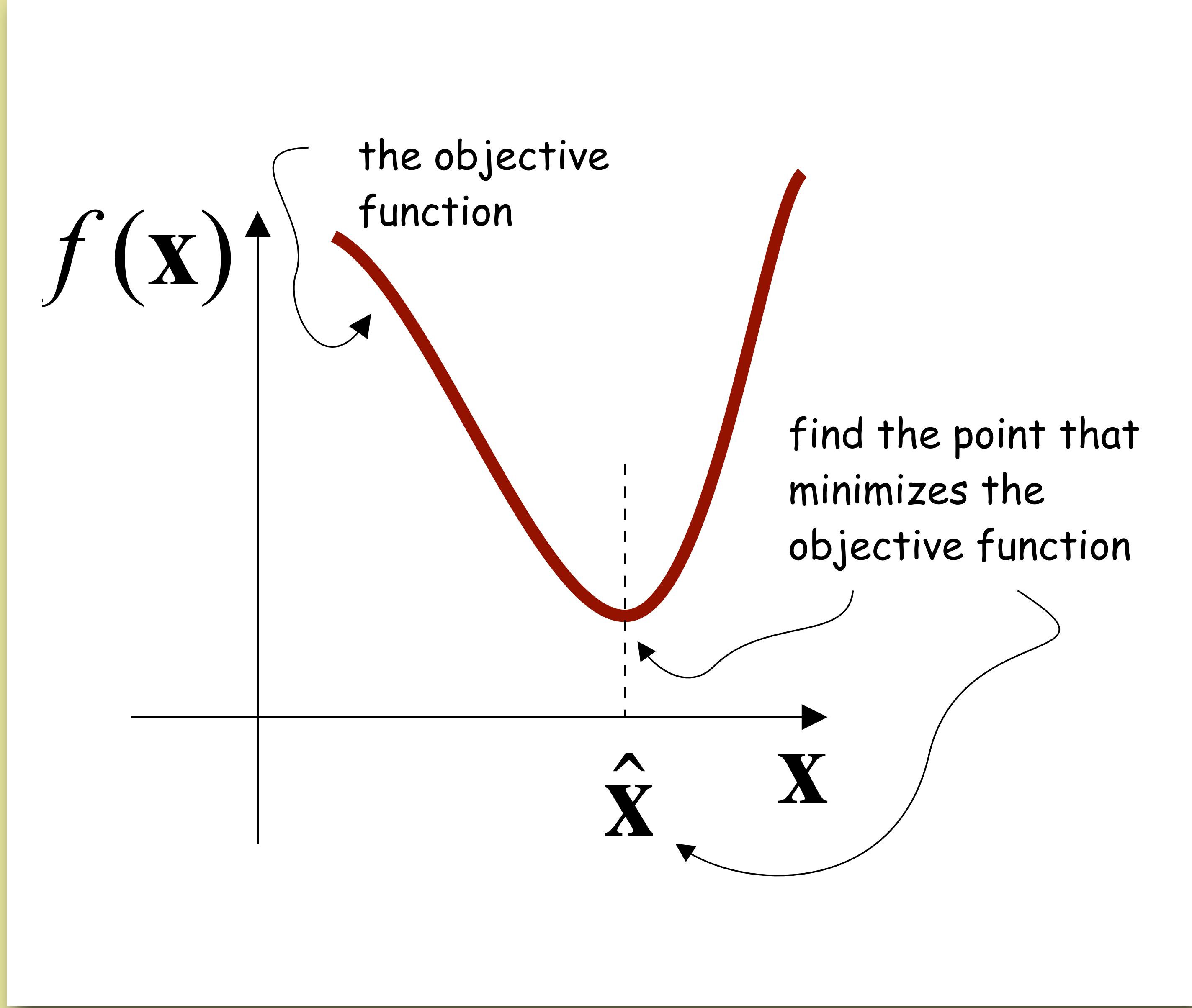
$$\hat{\mathbf{x}} = A^{-1}\mathbf{b}$$

the solution we seek

UNIVERSITY OF CALGARY

optimization

- as an alternative, can convert to an optimization problem
  - form *objective function*, $f(\mathbf{x})$
  - solve by finding minimum of $f(\mathbf{x})$



objective function

- here is an objective function that converts an n-by-n linear system into an objective function for optimization

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

from generic n-by-n system

UNIVERSITY OF CALGARY

some calculus

$$f(\mathbf{x}) = \frac{1}{2} \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} b_1 & b_2 & \cdots & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\frac{\partial f}{\partial x_1} = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n - b_1$$

UNIVERSITY OF CALGARY

gradient

- collect partial derivatives into a single equation

$$\begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} - \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$\nabla f = A\mathbf{x} - \mathbf{b}$$

the gradient of $f(\mathbf{x})$

UNIVERSITY OF CALGARY

optimize

- to find maximum or minimum set gradient to zero
- same as solving original n-by-n linear system

set the gradient to zero

$$\nabla f = 0 = A\mathbf{x} - \mathbf{b}$$

$$A\mathbf{x} = \mathbf{b}$$

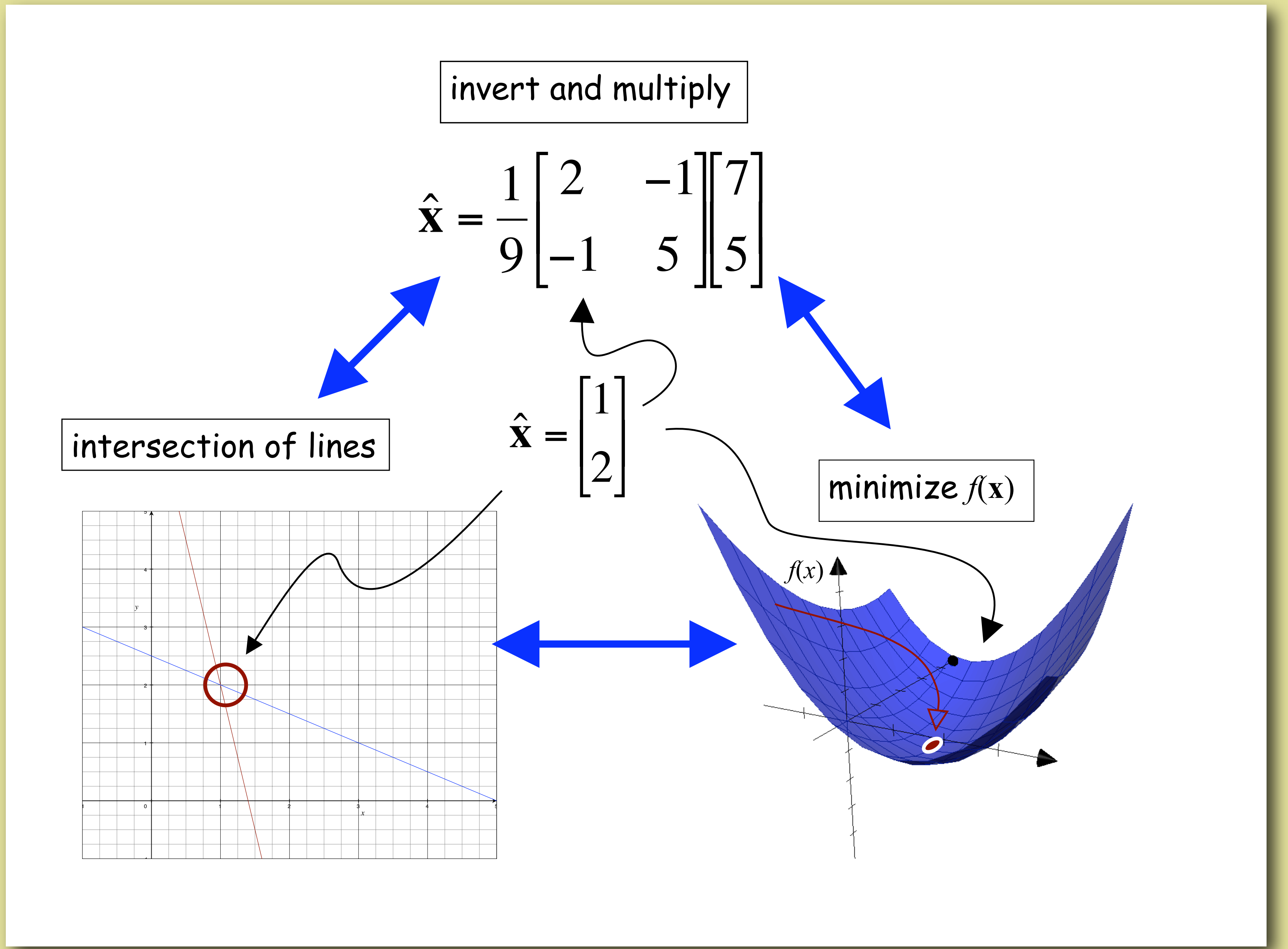
the generic n-by-n linear system

UNIVERSITY OF CALGARY

three ways to solve

$$A = \begin{bmatrix} 5 & 1 \\ 1 & 2 \end{bmatrix},$$

$$b = \begin{bmatrix} 7 \\ 5 \end{bmatrix}$$



why?

- faster
  - time to compute A^{-1} or *LU* decomposition is in $\Theta(n^3)$
  - often iterative optimization is faster
- give intuitive understanding of important properties of *A*

UNIVERSITY OF CALGARY

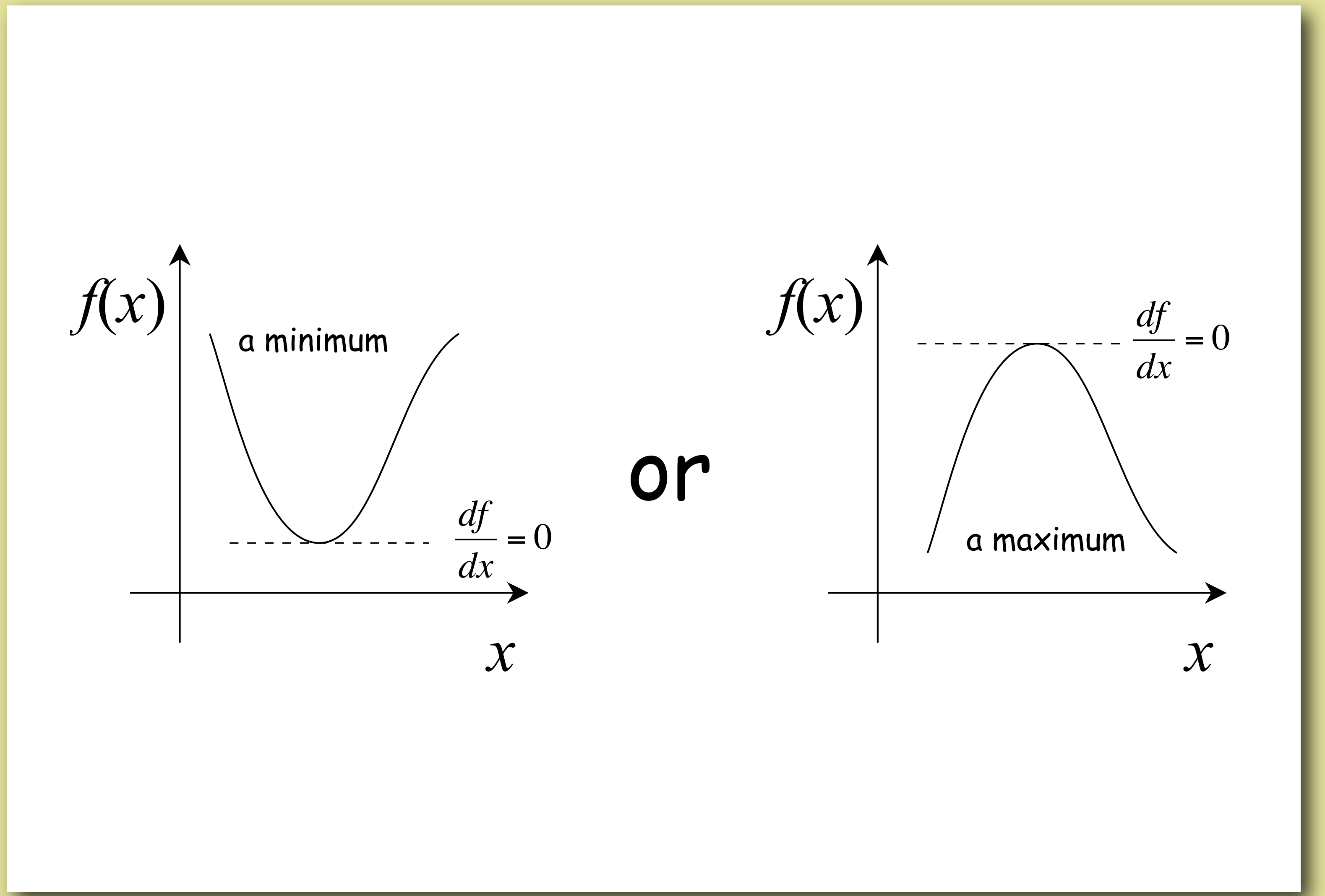
max, min, or ?

- in 1 dimension

$$\frac{d^2 f}{dx^2} > 0 \quad \text{minimum}$$

$$\frac{d^2 f}{dx^2} < 0 \quad \text{maximum}$$

$$\frac{d^2 f}{dx^2} = 0 \quad \text{inflection}$$



UNIVERSITY OF CALGARY

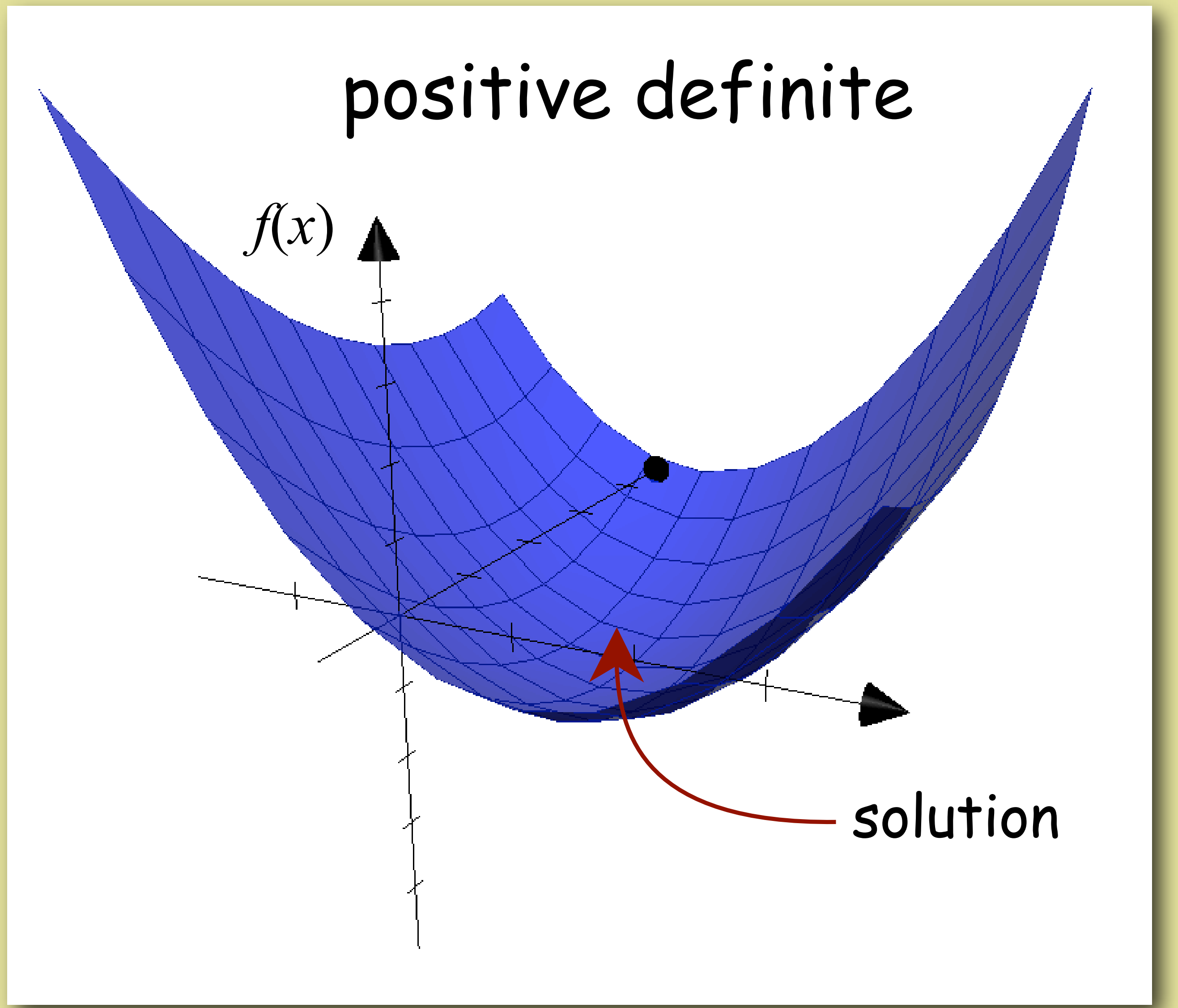
$n \geq 2$ is more complicated

positive definite

- more complicated with two or more dimension
- $f(\mathbf{x})$ has a minimum when A is positive definite

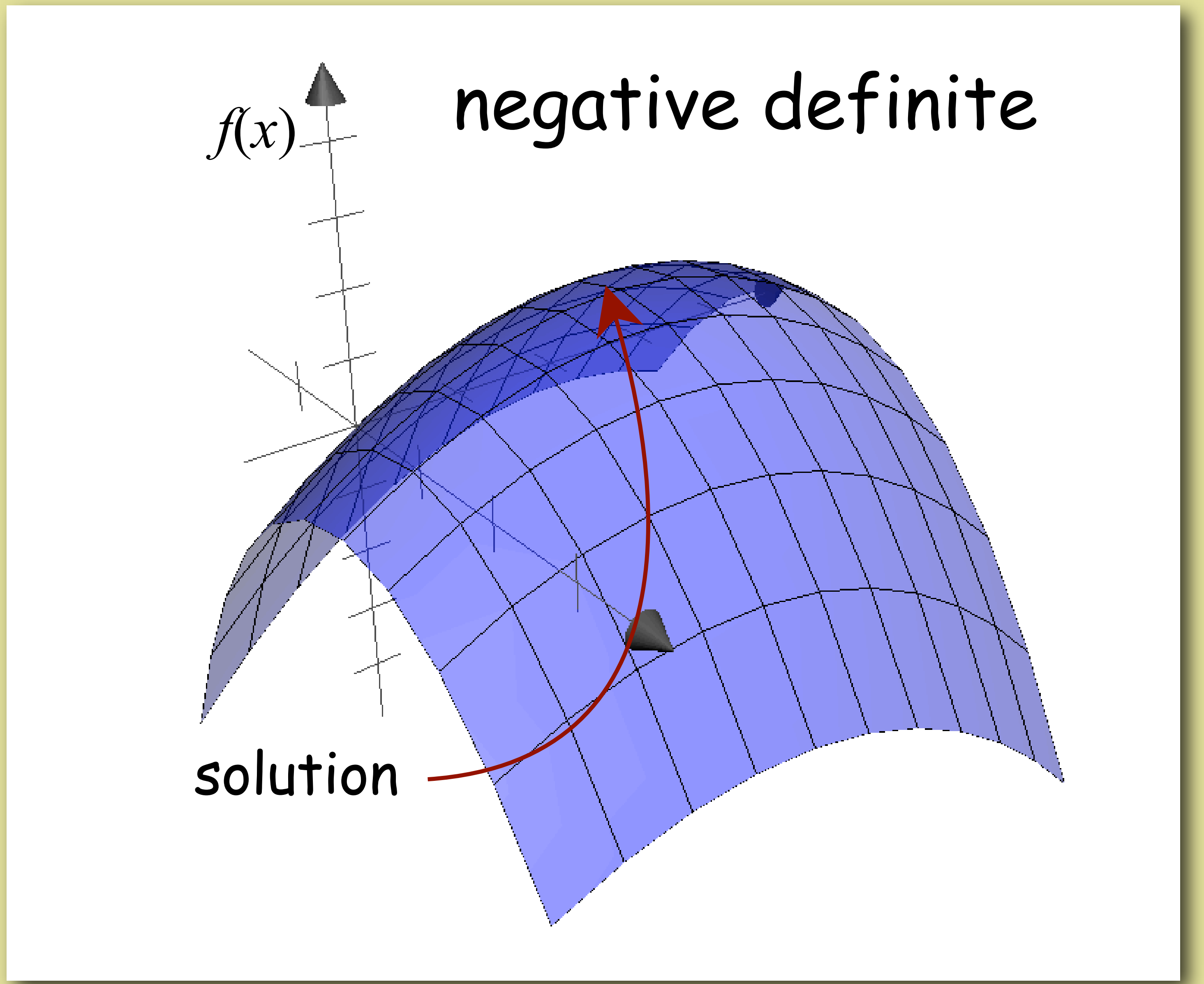
$$x^T A x \geq 0$$

- eigenvalues of A are all positive



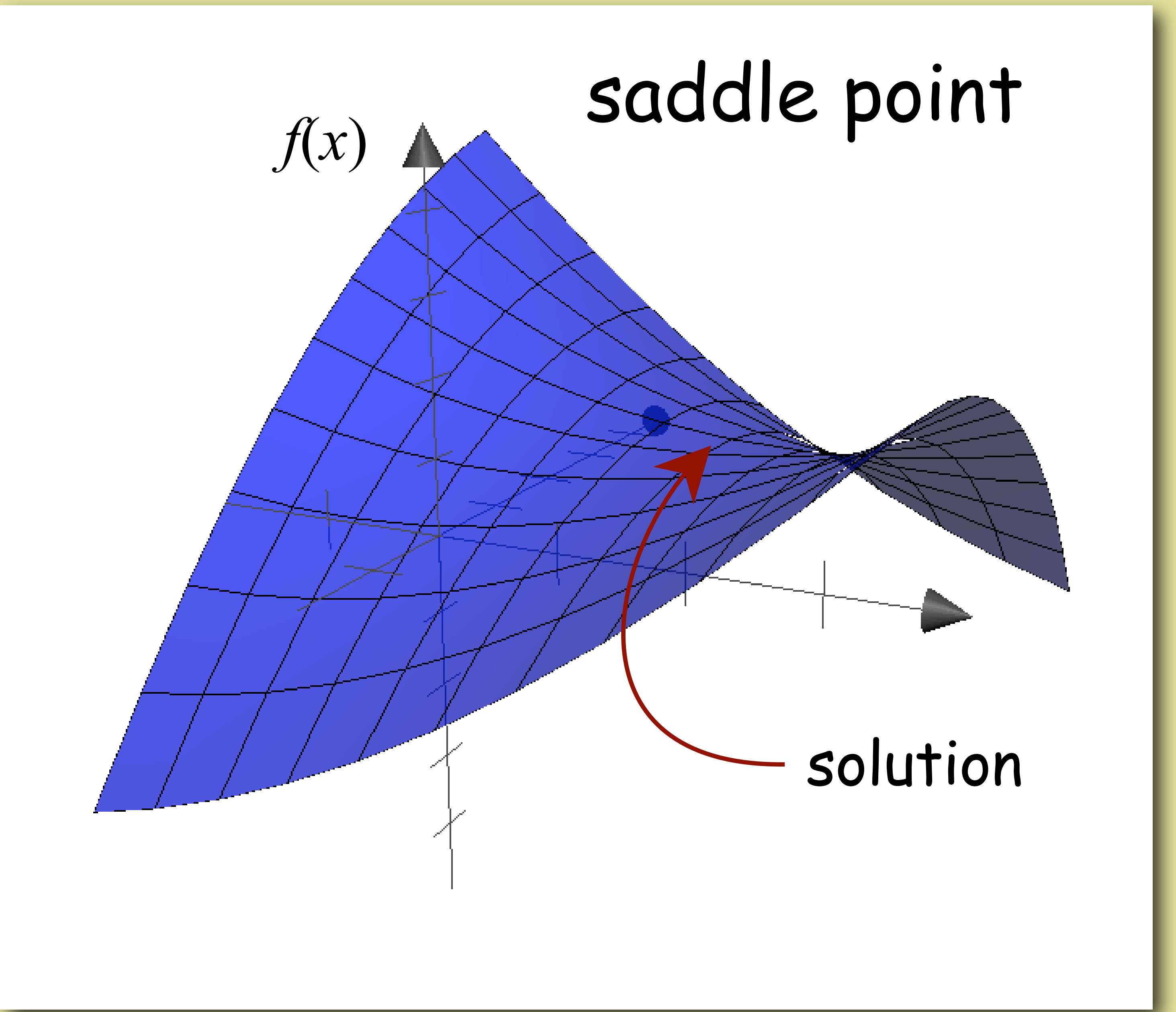
negative definite

- $f(\mathbf{x})$ has maximum when A is negative definite
- if A is negative definite all eigenvalues are negative
- can use optimization



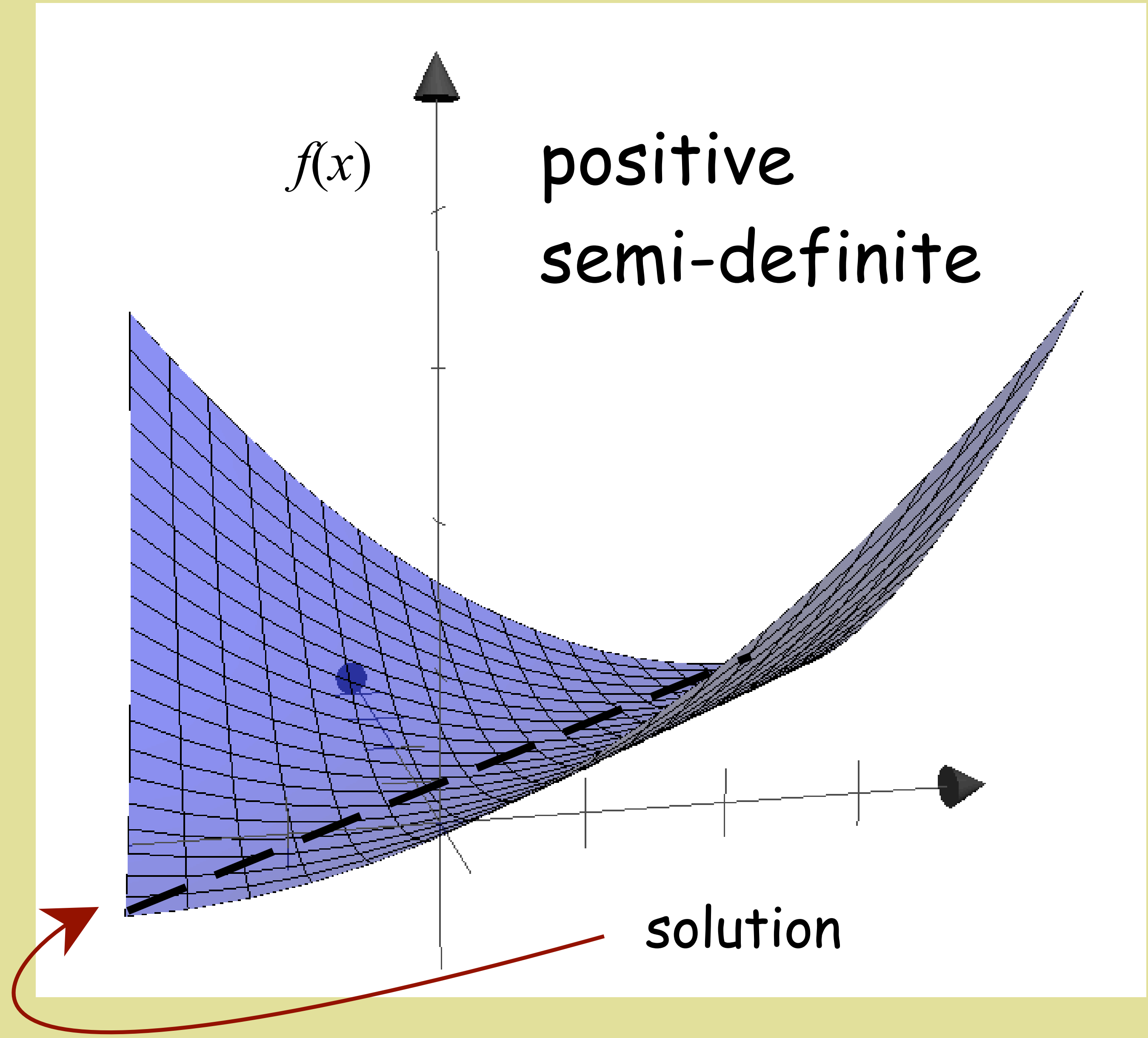
saddle point

- if neither positive definite or negative definite then
  - saddle point
- no max or min
  - cannot use optimization



semi-definite

- if some eigenvalues are greater than or equal to zero, then *A* is semi-definite
- minimum is not unique
- cannot minimize



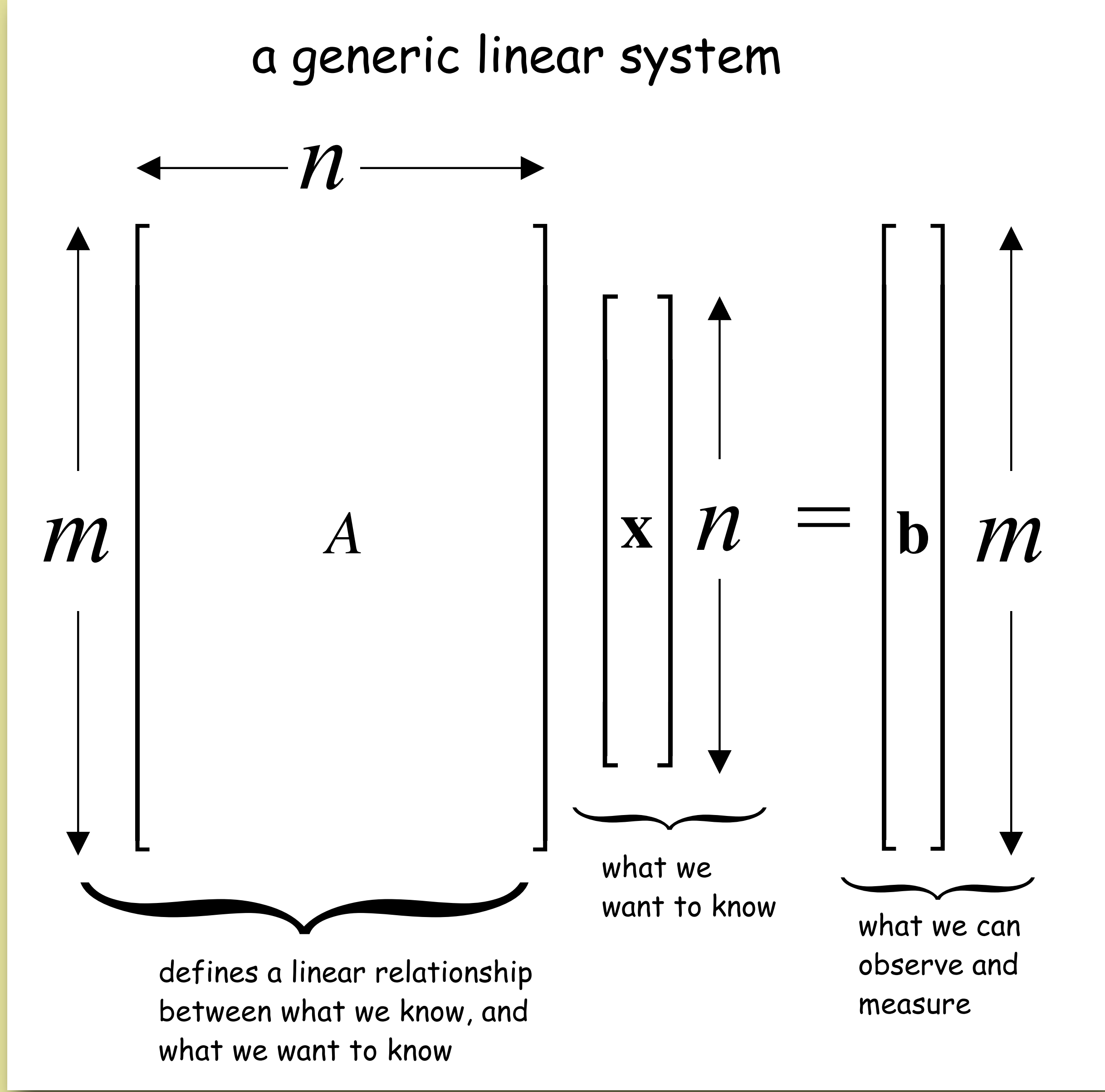
Least-Squares Solutions for Linear Systems

Jeffrey E. Boyd

UNIVERSITY OF CALGARY

generic linear system

- measure *m* values in **b**.
- want to know the *n* values in **x**.
- *A* defines relationship between **b** and **x**.



UNIVERSITY OF CALGARY

too many equations

- more equations than unkowns
- no solution
- find an *average* solution
- formulate as optimization
    - minimize *E*

$$
\begin{aligned}
E &= \|A\mathbf{x} - \mathbf{b}\|^2 \\
&= (A\mathbf{x} - \mathbf{b})^T (A\mathbf{x} - \mathbf{b}) \\
&= \mathbf{x}^T A^T A\mathbf{x} - \mathbf{b}^T A\mathbf{x} - \mathbf{x}^T A^T \mathbf{b} + \mathbf{b}^T \mathbf{b} \\
&= \mathbf{x}^T A^T A\mathbf{x} - 2\mathbf{b}^T A\mathbf{x} - \mathbf{b}^T \mathbf{b}
\end{aligned}
$$

UNIVERSITY OF CALGARY

compare

- compare to *f*(**x**)
- optimize *E* the same way we optimize *f*(**x**)
  - differentiate and set to zero

$$E = \mathbf{x}^T A^T A\mathbf{x} - 2\mathbf{b}^T A\mathbf{x} - \mathbf{b}^T\mathbf{b}$$

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

UNIVERSITY OF CALGARY

result

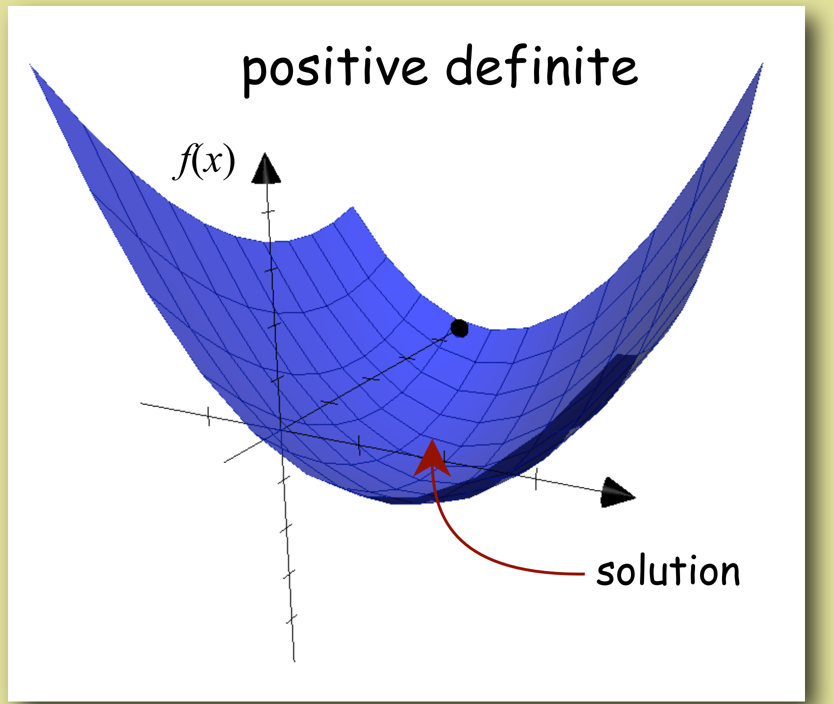
- optimize E by solving this linear system

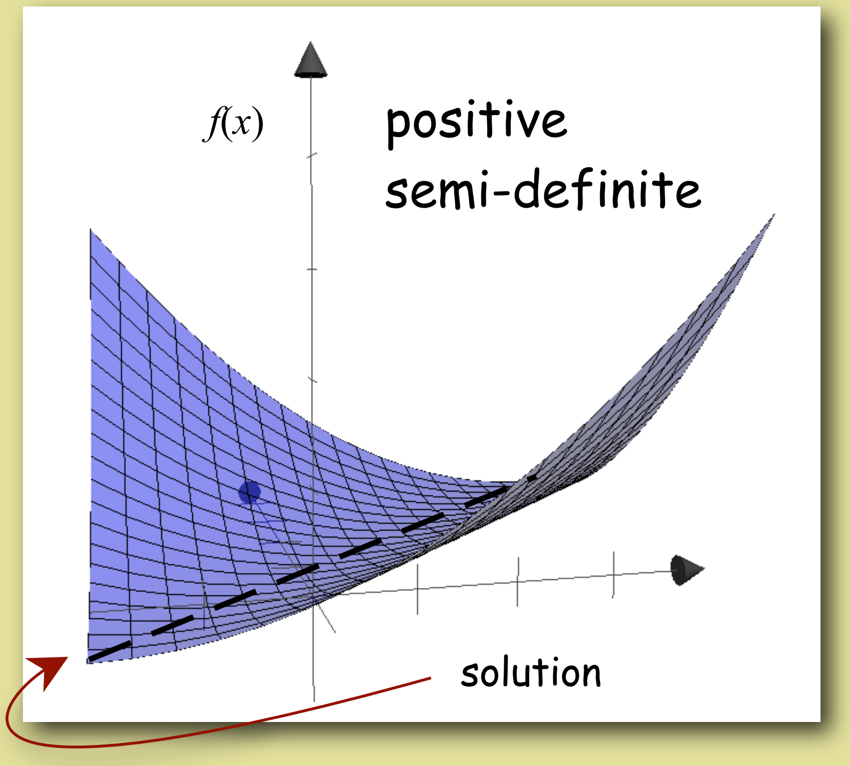
$$A^T A\hat{\mathbf{x}} = A^T\mathbf{b}$$

UNIVERSITY OF CALGARY

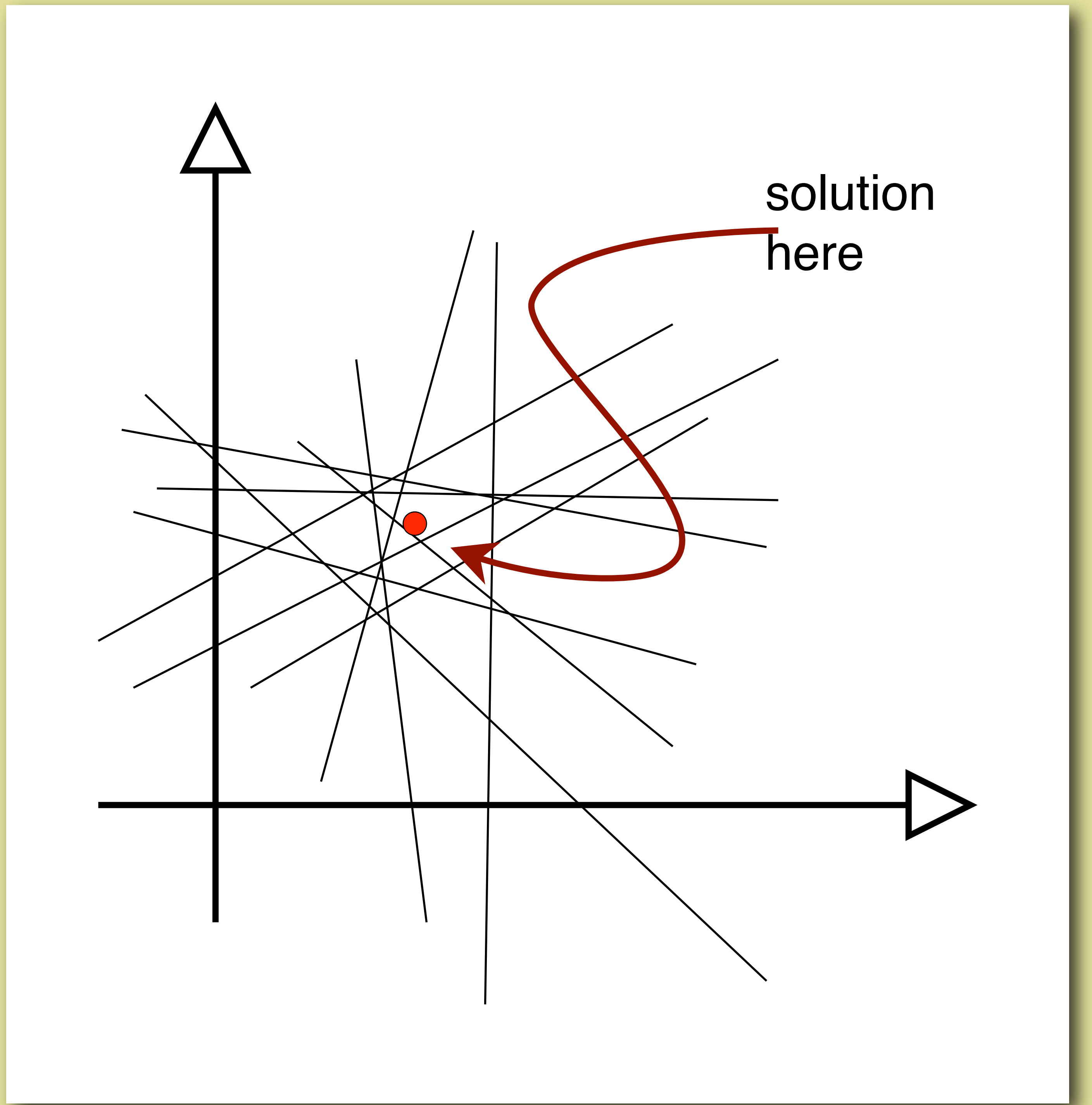
properties of A

- if columns of A independent
  - $m >= n$
  - no column linear combination of another
  - then A^TA is symmetric, positive definite
- if columns not independent
  - then A^TA is symmetric, positive semi-definite





graphically



sample average

- suppose I have a *m* measurements of the temperature in the room
- what is the temperature in the room?
- *x* is the temperature
- *b* is a measurement
- for each measurement I get one equation
  - $1x = b$

$$\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

UNIVERSITY OF CALGARY

LS average

- the least squares estimate of the temperature is the average

- this is a good way to understand LS intuitively

$$
\begin{aligned}
A &= \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \\
\mathbf{x} &= x \\
\mathbf{b} &= \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \\
A^T A\hat{\mathbf{x}} &= A^T \mathbf{b} \\
m\hat{x} &= \sum_{i=1}^{m} b_i \\
\hat{x} &= \frac{1}{m}\sum_{i=1}^{m} b_i
\end{aligned}
$$

UNIVERSITY OF CALGARY

weighted LS

- put more emphasis on some measurements than others with *W*.

$$
\begin{aligned}
WA\mathbf{x} &= W\mathbf{b} \\
A^T W^T W A\hat{\mathbf{x}} &= A^T W^T W\mathbf{b}
\end{aligned}
$$

W

- how do you choose *W*?
- for normally distributed measurement errors
  - $W^T W$ is the inverse of the C, covariance of measurement errors

$$
\begin{aligned}
\mathbf{e} &= A\mathbf{x} - \mathbf{b} \\
\sigma_{ij} &= \mathrm{cov}(e_i, e_j) \\
\sigma_{ii} &= \sigma_i^2 = \mathrm{cov}(e_i, e_i) \\
C &= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{n2} & \dots & \sigma_m^2 \end{bmatrix} \\
W^T W &= C^{-1}
\end{aligned}
$$

UNIVERSITY OF CALGARY

when *W* diagonal

- simplify for understanding
- consider when *W* is diagonal
  - measurement errors are independent
  - measurements with larger errors get less weight

$$
\begin{aligned}
WA\mathbf{x} &= W\mathbf{b} \\
\begin{bmatrix} \frac{1}{\sigma_1} & & & \\ & \frac{1}{\sigma_2} & & \\ & & \ddots & \\ & & & \frac{1}{\sigma_m} \end{bmatrix} A\mathbf{x} &= \begin{bmatrix} \frac{1}{\sigma_1} & & & \\ & \frac{1}{\sigma_2} & & \\ & & \ddots & \\ & & & \frac{1}{\sigma_m} \end{bmatrix} \mathbf{b} \\
W^T W = C^{-1} &= \mathrm{diag}(\frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2}, \ldots, \frac{1}{\sigma_m^2})
\end{aligned}
$$

UNIVERSITY OF CALGARY

computer vision

- consider
  - what you can measure
  - what you want to know
  - relationship between the two
- form linear system
- solve