COMPUTER SCIENCE

INDIANA UNIVERSITY

School of Informatics and Computing
Bloomington

Linear regression

Reminders

- Assignment #1 is due today Wednesday
- Next readings: Chapters 3 and 4
  - Chapter 3: Introduction to machine learning
  - Chapter 4: Linear regression
- Notation sheet
- We will review gradient descent and optimization

Thought questions

- Differences between variables and givens or constants

- Question: (Table 1.1 in page 31) has some "important" expectation functions. I see a distinction between first 6 rows and last three rows. The first three rows correspond to a function f: from R to R. E.g. 3rd row with k=2: f(x)=x^2. However, I don't see how one can represent the last 3 rows in the same template since the value of, say log(1/p_X(x)) depends not only on x but also p_X.

- p_X is just a function, which is some given

- We distinguish between the variable and the things held constant

Thought questions

$f(x)$	Symbol	Name
x	$E[X]$	Mean
$(x - E[X])^2$	$V[X]$	Variance
x^k	$E[X^k]$	k-th moment; $k \in \mathbb{N}$
$(x - E[X])^k$	$E[(X - E[X])^k]$	k-th central moment; $k \in \mathbb{N}$
e^{tx}	$M_X(t)$	Moment generating function
e^{itx}	$\varphi_X(t)$	Characteristic function
$\log \frac{1}{p_X(x)}$	$H(X)$	(Differential) entropy
$\log \frac{p_X(x)}{q(x)}$	$D(p_X\|\|q)$	Kullback-Leibler divergence
$\left(\frac{\partial}{\partial\theta} \log p_X(x\|\theta)\right)^2$	$\mathcal{I}(\theta)$	Fisher information

Examples

- When learning a parameter theta, the dataset composed of (xi,yi) is given and only theta is a variable (i.e., unknown)
  - Then we try to solve for theta
- When we are obtaining samples xi from a distribution, the distribution is constant/given, and the variable is the x that will be sampled (i.e., unknown until sampled)
- Distinction between components that are fixed/known and those that are variable/unknown
- When following or doing derivations, keep this in mind

Thought questions

- It’s stated that MAP and ML converge to the same solution for large data sets. Is there a difference in the rates of convergence between these two approaches, i.e. could it potentially be more useful to use MAP rather than ML despite being harder to calculate if MAP had a faster rate of convergence?
  - Imagine an oracle gave you the true parameter before learning
  - Intuitively it seems to make sense that if the prior chosen for MAP heavily biases the parameter to be close to the true parameter, then we should converge (get to the optimal solution) faster, since we started closer to it
  - If log prior is strongly convex, can improve convergence rate
  - Different types of convergence: convergence to solution for fixed # of samples and convergence to true parameters with infinite samples

Summary

- Expected cost introduced to formalize our objective

- Bayes risk function indicates best we could do
  - f(x) specified for each x, rather than having some simpler (continuous) function class
  - can think of it as a table of values

- For classification (with uniform cost)

$$f^*(\mathbf{x}) \quad = \arg\max_{y \in \mathcal{Y}} \{p(y|\mathbf{x})\} \, .$$

- For regression (with least-squares cost)

$$f^*(\mathbf{x}) = \int_{\mathcal{Y}} y p(y|\mathbf{x}) dy$$

Summary: Bayes optimal models

- In practice, cannot learn a “table of values” for a continuous space (i.e., the space of features x)

- Instead pick a function class $\mathcal{F}$

- Given a dataset $\mathcal{D}$

$$p(y|\mathbf{x}, \mathcal{D}) = \sum_{f \in \mathcal{F}} p(y|\mathbf{x}, f) p(f|\mathbf{x}, \mathcal{D}),$$

- For now, we will put all weight on one f in p(f I D): the MAP f

$$p(f|\mathcal{D}) = \begin{cases} 1 & \text{if } f \text{ is the MAP estimate for the data;} \\ 0 & \text{else.} \end{cases}$$

- Later we will return to other p(f I D)

Linear Regression

e.g.,
x_i = size of house
y_i = cost of house

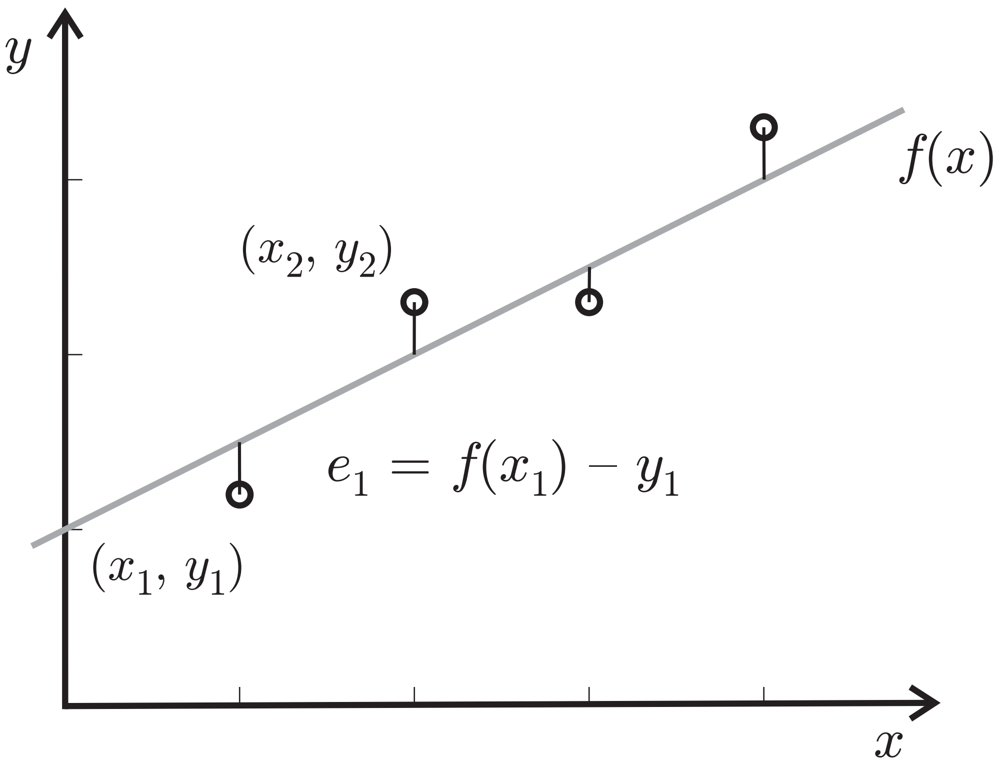
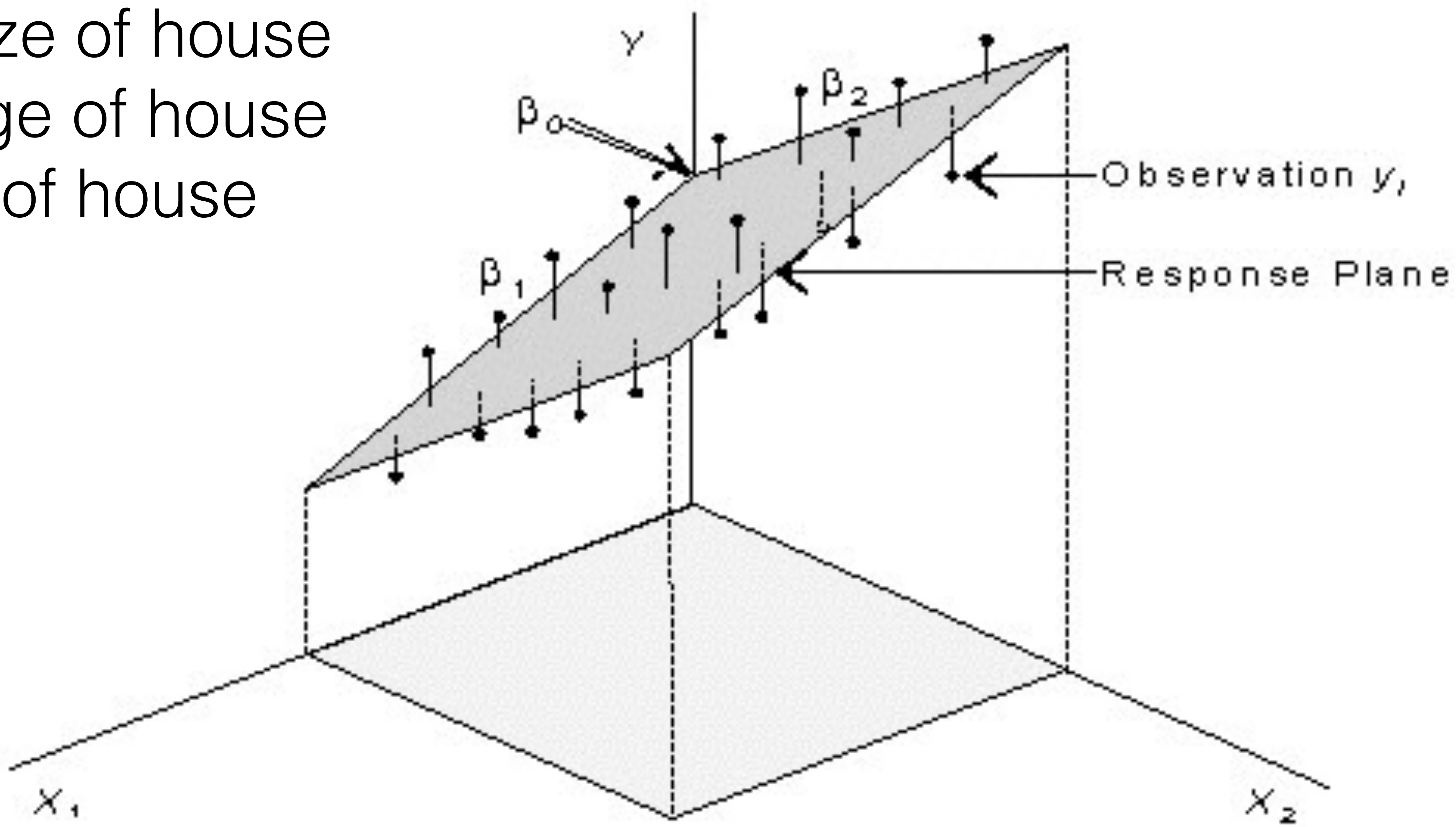


Figure 4.1: An example of a linear regression fitting on data set $\mathcal{D} = \{(1, 1.2), (2, 2.3), (3, 2.3), (4, 3.3)\}$. *The task of the optimization process is to find the best linear function* $f(x) = w_0 + w_1 x$ *so that the sum of squared errors* $e_1^2 + e_2^2 + e_3^2 + e_4^2$ *is minimized.*

(Multiple) Linear Regression

e.g.,
x_{i1} = size of house
x_{i2} = age of house
y_i = cost of house



What about multiple outputs y?

Linear regression importance

- Many other techniques will use linear weighting of features
  - including neural networks
- Often, we will add non-linearity using
  - non-linear transformations of linear weighting
  - non-linear transformations of features
- Becoming comfortable will linear weightings, for multiple inputs and outputs, is important

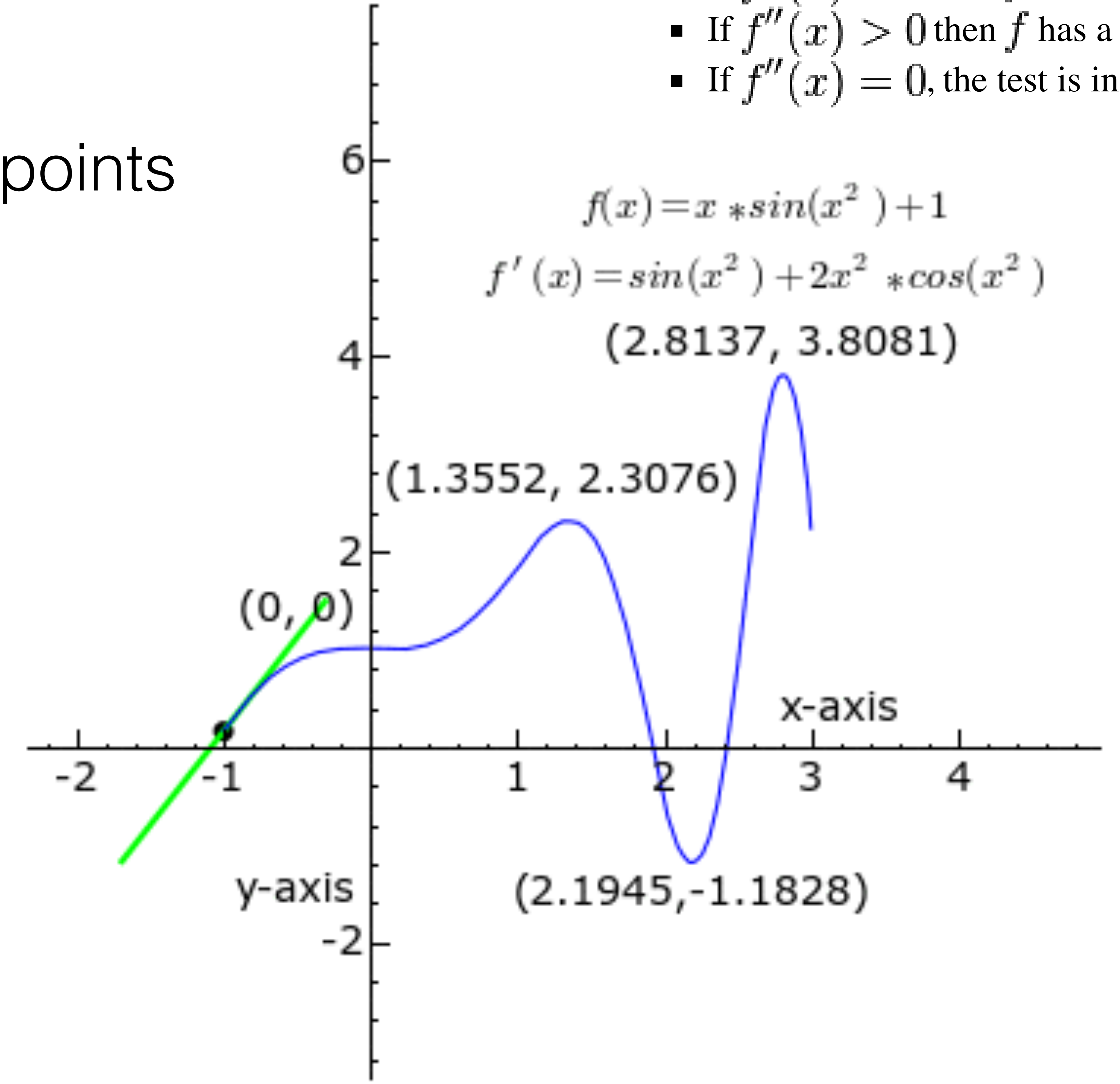
Whiteboard

- Maximum likelihood formulation (and assumptions)
- Solving the optimization
- Linear regression is a simple example of the more general topics we have been discussing so far

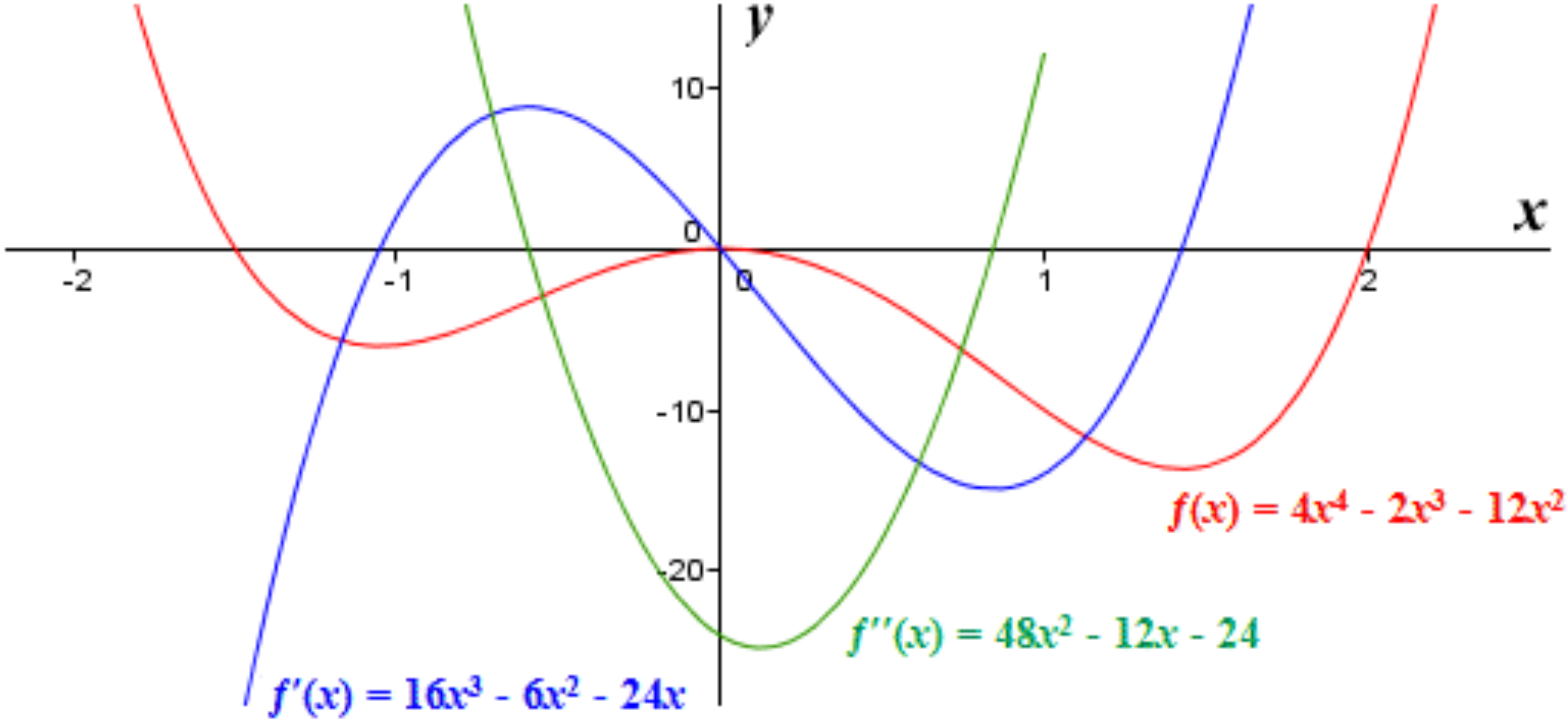
Quick calculus refresher

- Minima
- Maxima
- Saddle points

- If $f''(x) < 0$ then f has a local maximum at x.
- If $f''(x) > 0$ then f has a local minimum at x.
- If $f''(x) = 0$, the test is inconclusive.



Second derivative test



Hessian

- If H is positive definite at x, then local minimum at x
- If H is negative definite at x, then local maximum at x
- If H has both positive and negative eigenvalues at x, then a saddle point at x

$$\mathbf{H} = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2} & \dfrac{\partial^2 f}{\partial x_1\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_1\,\partial x_n} \\ \dfrac{\partial^2 f}{\partial x_2\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f}{\partial x_2\,\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n\,\partial x_1} & \dfrac{\partial^2 f}{\partial x_n\,\partial x_2} & \cdots & \dfrac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$