Optimization — Basics from Calculus

36-462/662, Spring 2022

1 February 2022 (Lecture 5)

Previously

- Risk of a strategy s is $r = \mathbb{E}[\ell(Y, s(X))]$, expected loss on new data
- Empirical risk of strategy s is $\hat{r}(s) = n^{-1} \sum_{i=1}^{n} \ell(Y_i, s(X_i))$, average loss on old data
- We want to find the empirical risk minimizer $\hat{s}$,

$$\hat{s} \equiv \operatorname*{argmin}_{s \in S} \hat{r}(s)$$

- We're now going to start opening up the black box of argmin

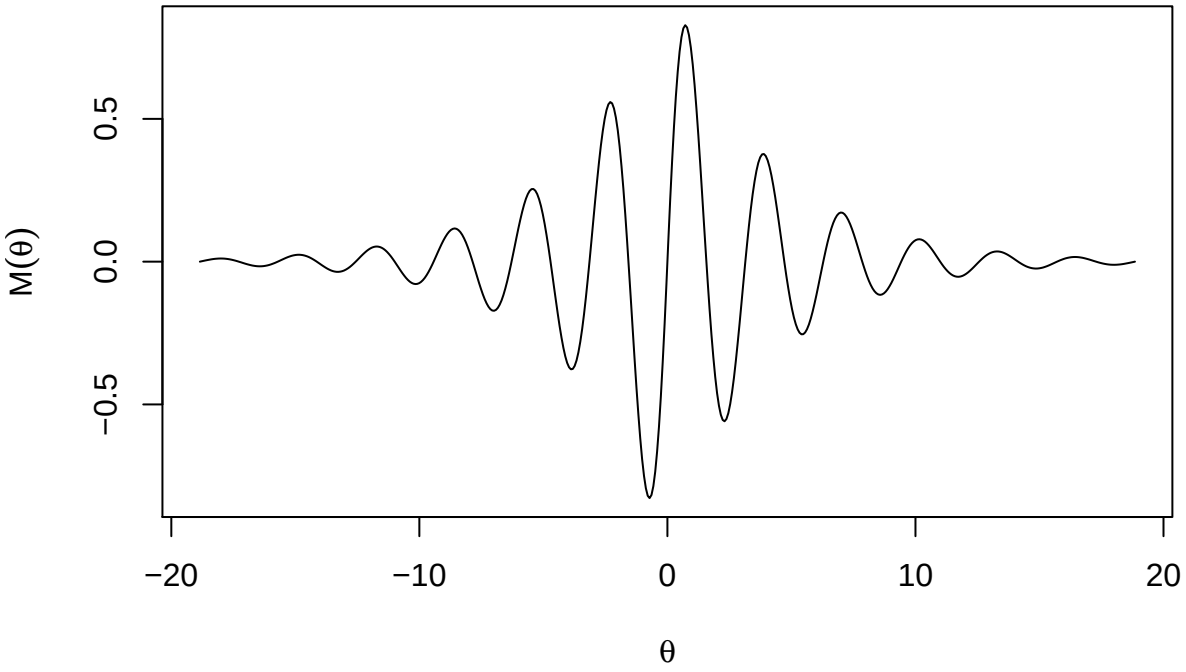
Optimization: some jargon

- The function we're trying to optimize is the **objective function**, let's say M today
- The argument to M is (say) θ
  - Some people call this the **optimand**
- The possible values of θ is Θ, the **domain** or **feasible set**, whose dimension is (say) p
- **Optimization** can be minimization or maximization, as we like; we'll stick with minimizing

Local vs. global minima

- θ is a **global minimum** when $\theta' \neq \theta \Rightarrow M(\theta') \geq M(\theta)$
  - Not necessarily unique!
- θ is a **local minimum** when $M(\theta) \leq M(\theta')$ whenever θ' is close enough to θ
  - Every global minimum is also a local minimum
  - If there's only one local minimum anywhere, it's the global minimum
- Lots of local minima tend to make it harder to find the global minimum

Local vs. global minima



"The" minimum: value vs. location

- If θ^* is a global minimum, then $M(\theta^*)$ is the **value** of the minimum or minimal value, in symbols

$$\min_{\theta \in \Theta} M(\theta)$$

- But θ^* itself is the **location** of the global minimum, in symbols

$$\operatorname*{argmin}_{\theta \in \Theta} M(\theta)$$

- Example: the minimal value of $(x-1)^2$ is 0, but the location of the minimum is $x = 1$
- Transformations: If L is an increasing function, then $L(M(\theta))$ has the same *location* for its minimum, but a different value
  - Example: log-likelihood vs. likelihood
- Both value and location can change with Θ
  - important later, when we look at constraints

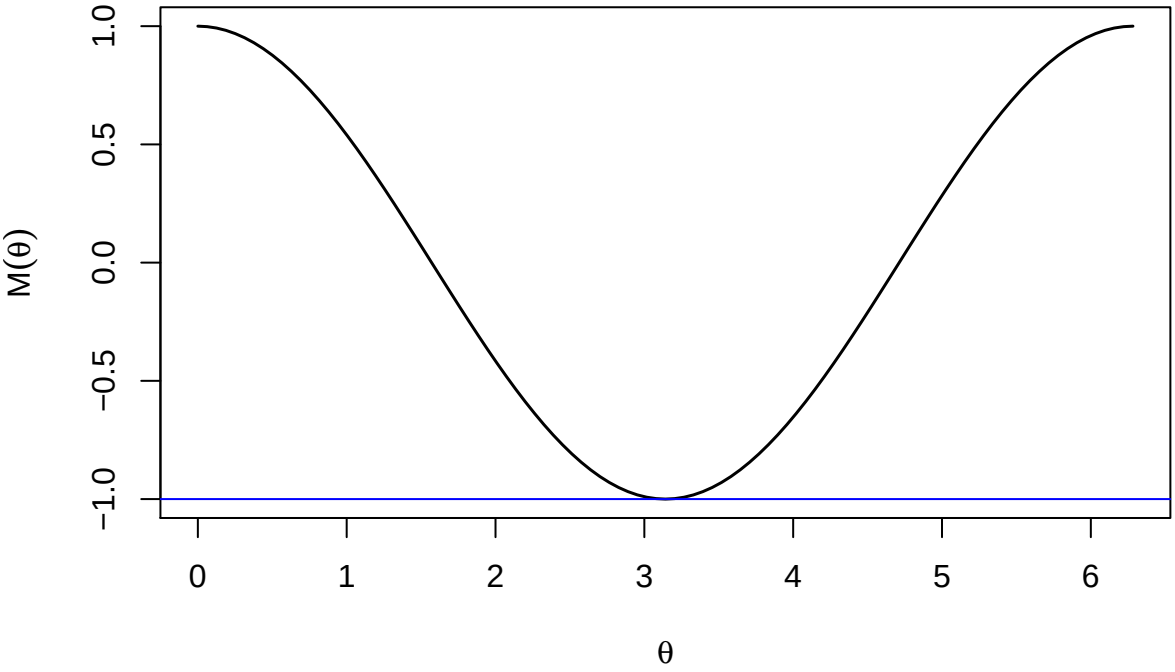
Finding the optimum: calculus basics

- Assume for now that θ is a continuous variable, and M is a nice, continuous function
  - We'll talk about not-so-nice situations later
- In fact, assume for now that θ is just a single real number
- Some things you probably remember from calculus about minima
  - Isn't $\frac{dM}{d\theta} = 0$?
  - Isn't $\frac{d^2M}{d\theta^2} > 0$?
- Yes, pretty much

The first order condition

- At an *interior*, minimum θ^*, $\frac{dM}{d\theta}(\theta^*) = 0$
  - If M had a slope, we could keep decreasing M by moving past θ^* in one direction or the other

The first order condition

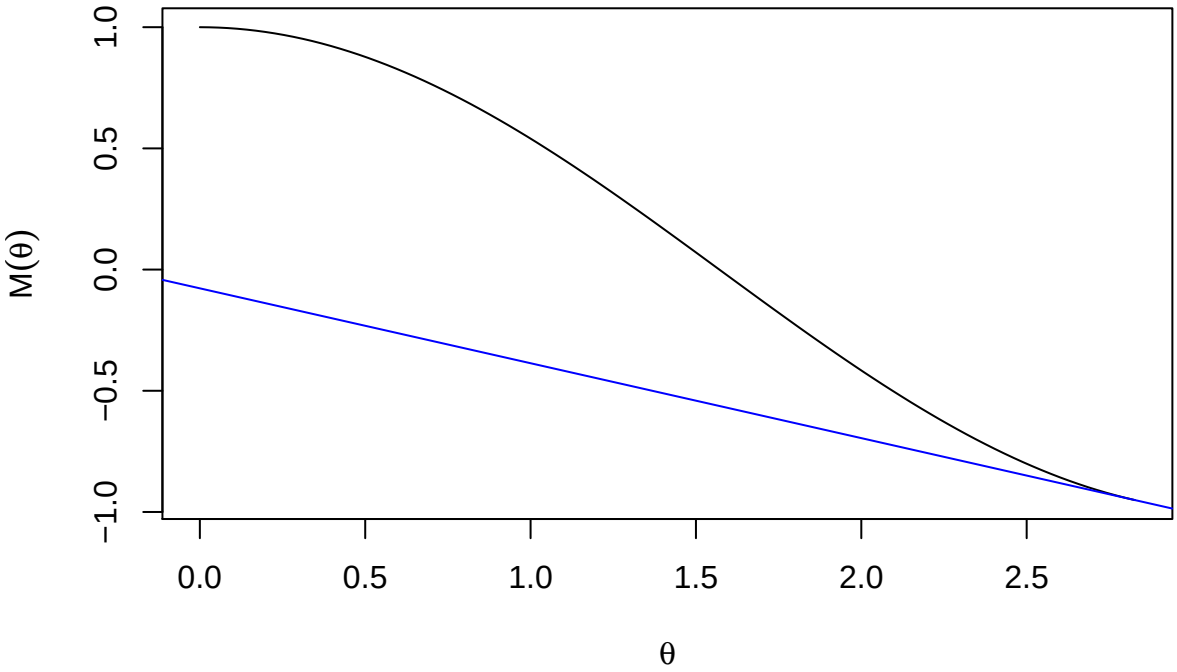


The tangent line to M is flat at the minimum θ^*

The first order condition and boundary optima

- At an *interior*, minimum θ^*, $\frac{dM}{d\theta}(\theta^*) = 0$
  - If M had a slope, we could keep decreasing M by moving past θ^* in one direction or the other
- This reasoning fails at the boundaries of Θ
  - Easy example: $\Theta = [0, 1]$, $M(\theta) = 1 - \theta$
  - Boundary optima *can* have zero slope though

The first order condition and boundary optima



The minimum on this domain is at the right-hand boundary, and the tangent line is not flat

The first order condition and boundary optima

- At an *interior*, minimum θ^*, $\frac{dM}{d\theta}(\theta^*) = 0$
  - If M had a slope, we could keep decreasing M by moving past θ^* in one direction or the other
- This reasoning fails at the boundaries of Θ
  - Easy example: $\Theta = [0, 1]$, $M(\theta) = 1 - \theta$
- But, except at boundaries, we need $\frac{dM}{d\theta}(\theta^*) = 0$
- This is called the **first-order** condition for a minimum

The second order condition

- Maxima as well as minima also have zero derivatives, so do inflection points
- A *sufficient* condition for a point with $dM/d\theta = 0$ to be a minimum: $d^2M/d\theta^2 > 0$
  - This is called the **second order** condition
  - Sufficient, but not necessary: θ^4 has a minimum at $\theta = 0$, even though $d^2M/d\theta^2 = 12\theta^2 = 0$ there
  - Minima which don't meet the second-order condition tend to be weird and fragile, like this
- Generally, we can find local minima in one dimension by using the first- and second- order conditions together:
  - Find all the solutions to $\frac{dM}{d\theta}(\theta^*) = 0$
  - Keep those with $\frac{d^2M}{d\theta^2}(\theta^*) > 0$

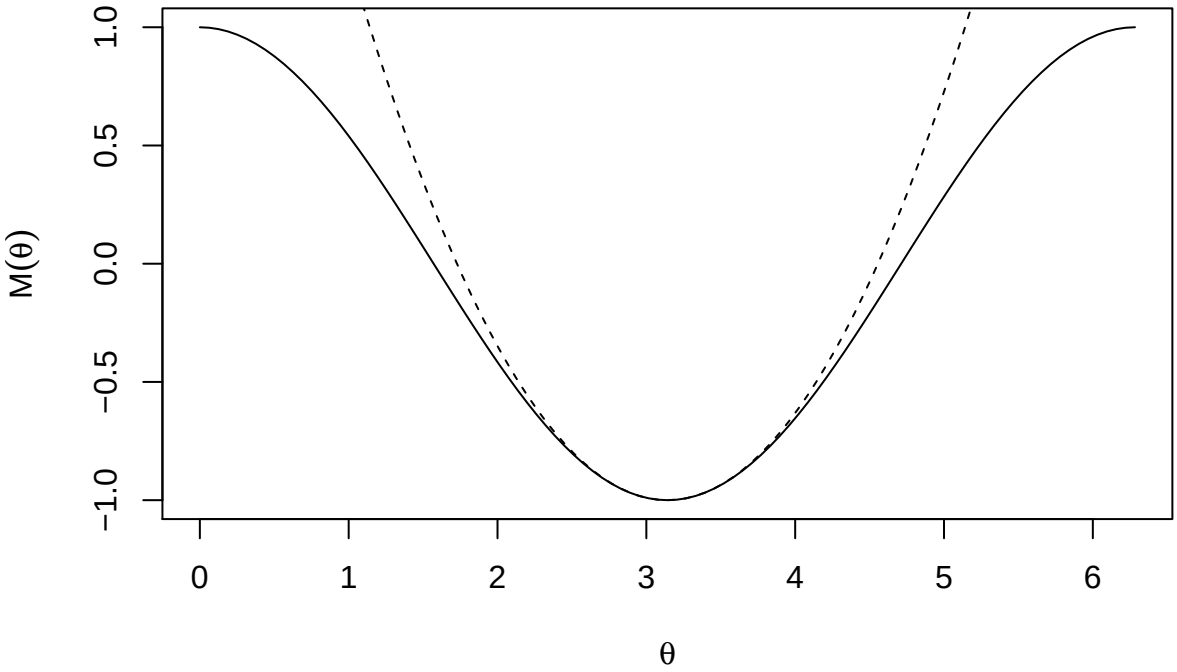
A bit more insight into the second-order condition

- Remember the definition of a derivative:

$$\frac{df}{dx}(x_0) \equiv \lim_{x \to x_0} f(x) - f(x_0)x - x_0$$

- Turn this around: for $x \approx x_0$,
$$f(x) \approx f(x_0) + (x - x_0)\frac{df}{dx}(x_0)$$
- This is a **first-order** Taylor approximation
- **Second-order** Taylor approximation: for $x \approx x_0$,
$$f(x) \approx f(x_0) + (x - x_0)\frac{df}{dx}(x_0) + \frac{1}{2}(x - x_0)^2\frac{d^2f}{dx^2}(x_0)$$
- First-order condition says: $\frac{dM}{d\theta}(\theta^*) = 0$
- So, near θ^*,
$$M(\theta) \approx M(\theta^*) + \frac{1}{2}(\theta - \theta^*)^2\frac{d^2M}{d\theta^2}(\theta^*)$$
- "Generic minima look, locally, like parabolas"

Generic minima look, locally, like parabolas



$M(\theta)$ (solid) vs. $M(\theta^*) + \frac{1}{2}(\theta - \theta^*)^2\frac{d^2M}{d\theta^2}(\theta^*)$ (dashed) around the local minimum θ^*

What about more than one dimension?

- Usually θ is a vector of $p > 1$ dimensions
- We can't, usually, do a separate optimization on each dimension
- What should happen at an interior minimum θ^*?
- M should have *no slope* at θ^* in *every* direction
  - Otherwise, we could lower the value of the function by moving
- M should increase as we move away from θ^* in every direction

No slope in any direction: the first-order condition

- Pick your favorite direction $\vec{v}$, a vector of length 1, say $(v_1, v_2, \ldots v_p)$
- The slope of M in that direction, at θ, is (chain rule)

$$\sum_{i=1}^{p} v_i \frac{\partial M}{\partial \theta_i}(\theta) = \vec{v} \cdot \nabla M(\theta)$$

- Here $\nabla M(\theta)$ is the **gradient** of M at θ, the vector of partial derivatives

$$\nabla M(\theta) = \left[\begin{array}{ccc} \frac{\partial M}{\partial \theta_1}(\theta) & \ldots & \frac{\partial M}{\partial \theta_p}(\theta) \end{array} \right]$$

  - No slope in *any* direction at θ^* means: $\vec{v} \cdot \nabla M(\theta^*) = 0$ for all $\vec{v} \neq 0$
  - And *that* means: $\nabla M(\theta^*) = 0$
  - The **first-order condition** is: "the gradient vanishes at the optimum"

First-order condition or first-order *conditions*?

- We have one vector equation $\nabla M(\theta^*) = 0$
- This is the same as a *system* of p equations for the partial derivatives:

$$\begin{aligned} \frac{\partial M}{\partial \theta_1}(\theta^*) &= 0 \\ &\vdots \\ \frac{\partial M}{\partial \theta_p}(\theta^*) &= 0 \end{aligned}$$

- This is good because we also have p unknowns, $\theta^* = \left[\begin{array}{ccc} \theta_1^* & \ldots & \theta_p^* \end{array} \right]$
- p equations for p unknowns $\Rightarrow$ typically a solution
  - Typically a *unique* solution if all the equations are linear in θ^*
  - Often not unique because nonlinear in θ^*
  - But still, there are solutions!

The function increases in every direction: the second-order condition

- Second-order Taylor series for vectors:

$$M(\theta) \approx M(\theta^*) + (\theta - \theta^*) \cdot \nabla M(\theta^*) + \frac{1}{2}(\theta - \theta^*) \cdot (\nabla \nabla M(\theta^*))\,(\theta - \theta^*)$$

- Here $\nabla \nabla M(\theta^*)$ is the matrix of second partial derivatives, $\frac{\partial^2 M}{\partial \theta_i \partial \theta_j}$, a.k.a. the **Hessian**
- First-order condition says the gradient term is zero at θ^*, so

$$M(\theta) \approx M(\theta^*) + \frac{1}{2}(\theta - \theta^*) \cdot (\nabla \nabla M(\theta^*))\,(\theta - \theta^*)$$

- θ^* is a minimum means:

$$(\theta - \theta^*) \cdot (\nabla \nabla M(\theta^*))\,(\theta - \theta^*) > 0$$

Positive-definite matrices

- A square matrix $\mathbf{h}$ is **positive-definite** when, for any non-zero vector $\vec{v}$,

$$\vec{v} \cdot \mathbf{h}\vec{v} > 0$$

  - If we only have $\vec{v} \cdot \mathbf{h}\vec{v} \geq 0$ then $\mathbf{h}$ is only **non-negative-definite** (or **positive semi-definite**)
- *Not* the same as $\mathbf{h}$ only having positive entries!
  - E.g., $\mathbf{p} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ is positive-definite
  - E.g., $\mathbf{n} = \begin{bmatrix} 0.5 & 1 \\ 1 & 0.5 \end{bmatrix}$ is not positive-definite
- We write this as $\mathbf{h} \succ 0$
  - Non-negative-definite is $\mathbf{h} \succeq 0$
- For symmetric matrices: $\mathbf{h}$ is positive definite $\Leftrightarrow$ all eigenvalues of $\mathbf{h}$ are > 0
  - The Hessian matrix $\nabla\nabla M$ is always symmetric (why?)
  - We'll do a refresher on eigenvalues in a few weeks before we really need them

The first- and second- order conditions for minima

For θ^* to be a local minimum,

- First-order condition: "The gradient must vanish", $\nabla M(\theta^*) = 0$
  - Necessary, except at a boundary
- Second-order condition: "The Hessian should be positive-definite", $\nabla\nabla M(\theta^*) \succ 0$
  - Sufficient; minima where it's violated are weird and a-typical

Near a minimum, nice functions look quadratic

- Go back to the Taylor approximation: if θ^* is a local minimum, so $\nabla M(\theta^*) = 0$, then

$$M(\theta) \approx M(\theta^*) + \frac{1}{2}(\theta - \theta^*) \cdot (\nabla\nabla M(\theta^*))\,(\theta - \theta^*)$$

- Consequence: if we come *close* to the minimum, so $\|\theta - \theta^*\| = \epsilon \ll 1$, then

$$M(\theta) \approx M(\theta^*) + O(\epsilon^2)$$

- If we can get ϵ-close to the *location* of the optimum, we get $O(\epsilon^2)$-close to the *value* of the optimum (and $\epsilon^2 \ll \epsilon \ll 1$)
  - Turned around, to get within δ of the *value* of the optimum, we need to only get with $O(\sqrt{\delta})$ of the *location* (and $\delta \ll \sqrt{\delta} \ll 1$)

Minimizing risk vs. minimizing empirical risk

- We *want* to minimize risk,

$$\theta^* = \operatorname*{argmin}_{\theta \in \Theta} r(\theta) = \operatorname*{argmin}_{\theta \in \Theta} \mathbb{E}\left[\ell(Y, s(X))\right]$$

- We *can* minimize empirical risk,

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \hat{r}(\theta) = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, s(x_i))$$

- We're going to see later that

$$\|\widehat{\theta} - \theta^*\| = O(1/\sqrt{n})$$

  - Basically: because of the law of large numbers
  - Assuming θ has finite dimensions which don't change with n

- Consequence:
$$r(\widehat{\theta}) \approx r(\theta^*) + O(1/n)$$
with factors from the Hessian buried inside the big O
- $\Rightarrow$ Minimizing the empirical risk comes closer and closer to minimizing the true risk

Morals to remember, about minimizing smooth functions

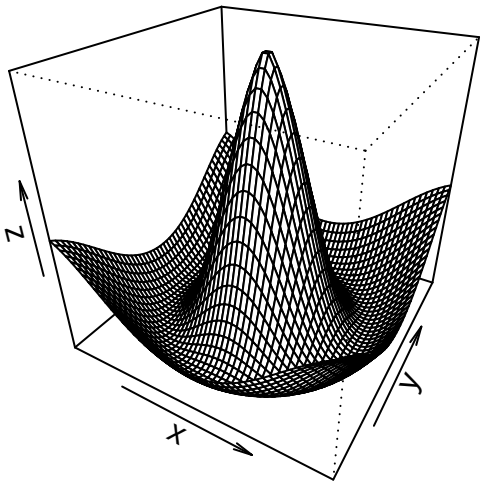
- Local vs. global minima
- First-order condition: "the gradient vanishes", $\nabla M(\theta^*) = 0$
  - Except at boundaries
- Second-order condition: "the Hessian is positive-definite", $\nabla\nabla M(\theta^*) \succ 0$
  - Except for weird, a-typical situations
- "Near a minimum, nice functions look quadratic"
- $\Rightarrow$ Coming within $O(\epsilon)$ of the *location* of the minimum puts us within $O(\epsilon^2)$ of the *value* of the minimum

Next time: actual algorithms

- How do we get the computer to actually *use* all this calculus?
  - Algorithms for optimization based on these and related ideas
- What happens because the computer can't do calculus *exactly*?
  - Optimization error and its consequences

Backup: What if $\nabla\nabla M \succeq 0$?

- What if the Hessian is only non-negative-definite, or positive-semi-definite?
- Then there's (at least) one direction $\vec{v}$ where
$$\vec{v} \cdot \nabla\nabla M \vec{v} = 0$$
- This suggests that if we start at θ^* and take a *small enough* step in the direction $\vec{v}$, we don't (necessarily) increase M
- We can have this when there is a *continuous set* of minima
  - Imagine a bowl where the base is raised in the middle — there's a ring of minima around the center
- This is a weird and delicate situation



Backup: Big-O notation

- $f(x) = O(g(x))$ as $x \to \infty$ means: there's some $C > 0$ so $|f(x)| \leq Cg(x)$ for all sufficiently big x
  - E.g., $10000000 + e^{-x} = \mathrm{O}(1)\$$
  - E.g., $37x^2 + 42x + 1421 = O(x^2)$
  - "Is at most of the order of", sometimes abbreviated "is of the order of"
  - For relevance, typically try to give the *tightest* bound we can, $37x^2 = O(x^4)$ but that's not *informative*
  - Use the same notation for limits $x \to 0$
- Small o notation: $f(x) = o(g(x))$ means: $\lim \frac{f(x)}{g(x)} = 0$

Backup: What do I mean when I say "weird, a-typical"?

- The set D is **dense** in another set A when there's a point in D arbitrarily close to every point in A
  - E.g., the rationals are dense in $[0, 1]$
- The set N is **nowhere dense** in A when it's not dense in any open subset of A
  - Open intervals: think of say $(1/4, 3/4)$, as opposed to $[1/4, 3/4]$
  - On the line, open sets are, roughly, unions of a countable number of open intervals; similarly in $\mathbb{R}^d$
- The set M is **meager** if it's a countable union of nowhere-dense sets
  - The rational numbers are meager, because there's only (!) a countable infinity of them, and each of them is nowhere-dense
- A set is **typical** if its complement is meager
  - Alternately: a set is typical if it's both open and dense
  - The irrational numbers in $[0, 1]$ are typical
- Local minima of smooth functions with positive second derivatives are typical, those with zero second derivatives are not typical
  - If you start from a minimum which *does* have a positive second derivative, you can continuously adjust it by arbitrarily small amounts and it still has a minimum at the same location with a positive second derivative (set is open and sense)
  - If you find a function with a zero second derivative, there are arbitrarily small tweaks to the function where you now have the same minimum but a positive second derivative
    - e.g., x^4 vs $x^4 + \epsilon x^2$, for $\epsilon > 0$ as small as you like
- These notions come from topology, which started by asking what properties of shapes stay the same under smooth transformations

References