TOPIC 0

Introduction

1 Review

Course: Math 535 (http://www.math.wisc.edu/~roch/mmids/) - Mathematical Methods in Data Science (MMiDS)
Author: Sebastien Roch (http://www.math.wisc.edu/~roch/), Department of Mathematics, University of Wisconsin-Madison
Updated: Sep 21, 2020

We first review a few basic concepts.

1.1 Vectors and norms

For a vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d$$

the Euclidean norm of $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^d x_i^2} = \sqrt{\mathbf{x}^T \mathbf{x}} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$$

where $\mathbf{x}^T$ denotes the transpose of $\mathbf{x}$ (seen as a single-column matrix) and

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{i=1}^d u_i v_i$$

is the inner product (https://en.wikipedia.org/wiki/Inner_product_space) of $\mathbf{u}$ and $\mathbf{v}$. This is also known as the 2-norm.

More generally, for $p \geq 1$, the p-norm (https://en.wikipedia.org/wiki/Lp_space#The_p-norm_in_countably_infinite_dimensions_and_ℓ_p_spaces) of $\mathbf{x}$ is given by

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}.$$

Here (https://commons.wikimedia.org/wiki/File:Lp_space_animation.gif#/media/File:Lp_space_animation.gif) is a nice visualization of the unit ball, that is, the set $\{\mathbf{x} : \|x\|_p \leq 1\}$, under varying p.

There exist many more norms. Formally:

Definition (Norm): A norm is a function ℓ from $\mathbb{R}^d$ to $\mathbb{R}_+$ that satisfies for all $a \in \mathbb{R}$, $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$

- *(Homogeneity):* $\ell(a\mathbf{u}) = |a|\ell(\mathbf{u})$
- *(Triangle inequality):* $\ell(\mathbf{u} + \mathbf{v}) \leq \ell(\mathbf{u}) + \ell(\mathbf{v})$
- *(Point-separating):* $\ell(u) = 0$ implies $\mathbf{u} = 0$.

◁

The triangle inequality for the 2-norm follows (https://en.wikipedia.org/wiki/Cauchy–Schwarz_inequality#Analysis) from the Cauchy–Schwarz inequality (https://en.wikipedia.org/wiki/Cauchy–Schwarz_inequality).

Theorem (Cauchy–Schwarz): For all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$

$$|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_2.$$

The Euclidean distance (https://en.wikipedia.org/wiki/Euclidean_distance) between two vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^d$ is the 2-norm of their difference

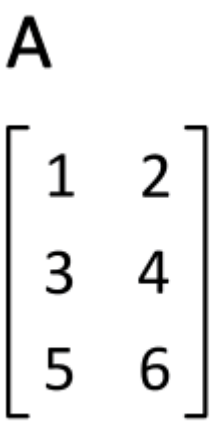
$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2.$$

Throughout we use the notation $\|\mathbf{x}\| = \|\mathbf{x}\|_2$ to indicate the 2-norm of $\mathbf{x}$ unless specified otherwise.

We will often work with collections of n vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^d$ and it will be convenient to stack them up into a matrix

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{bmatrix},$$

where ^{T} indicates the transpose (https://en.wikipedia.org/wiki/Transpose):



(Source (https://commons.wikimedia.org/wiki/File:Matrix_transpose.gif))

NUMERICAL CORNER In Julia, a vector can be obtained in different ways. The following method gives a row vector as a two-dimensional array.

In [1]:
```
# Julia version: 1.5.1
using LinearAlgebra, Statistics, Plots
```

In [2]:
```
t = [1. 3. 5.]
```

Out[2]:
```
1×3 Array{Float64,2}:
 1.0  3.0  5.0
```

To turn it into a one-dimensional array, use `vec` (https://docs.julialang.org/en/v1/base/arrays/#Base.vec).

In [3]:
```
vec(t)
```

Out[3]:
```
3-element Array{Float64,1}:
 1.0
 3.0
 5.0
```

To construct a one-dimensional array directly, use commas to separate the entries.

In [4]:
```
u = [1., 3., 5., 7.]
```

Out[4]:
```
4-element Array{Float64,1}:
 1.0
 3.0
 5.0
 7.0
```

To obtain the norm of a vector, we can use the function `norm` (https://docs.julialang.org/en/v1/stdlib/LinearAlgebra/#LinearAlgebra.norm) (which requires the `LinearAlgebra` package):

In [5]:
```
norm(u)
```

Out[5]:
```
9.16515138991168
```

which we can check "by hand"

In [6]:
```
sqrt(sum(u.^2))
```

Out[6]:
```
9.16515138991168
```

The `.` above is called broadcasting (https://docs.julialang.org/en/v1.2/manual/mathematical-operations/#man-dot-operators-1). It applies the operator following it (in this case taking a square) element-wise.

Exercise: Compute the inner product of $u = (1, 2, 3, 4)$ and $v = (5, 4, 3, 2)$ without using the function `dot` (https://docs.julialang.org/en/v1/stdlib/LinearAlgebra/#LinearAlgebra.dot). Hint: The product of two real numbers a and b is `a * b`.

In [7]:
```
# Try it!
u = [1., 2., 3., 4.];
# EDIT THIS LINE: define v
# EDIT THIS LINE: compute the inner product between u and v
```

◁

To create a matrix out of two vectors, we use the function `hcat` (https://docs.julialang.org/en/v1.2/base/arrays/#Base.hcat) and transpose.

In [8]:
```
u = [1., 3., 5., 7.];
v = [2., 4., 6., 8.];
X = hcat(u,v)'
```

Out[8]:
```
2×4 Adjoint{Float64,Array{Float64,2}}:
 1.0  3.0  5.0  7.0
 2.0  4.0  6.0  8.0
```

With more than two vectors, we can use the `reduce` (https://docs.julialang.org/en/v1/base/collections/#Base.reduce-Tuple{Any,Any}) function.

In [9]:
```
u = [1., 3., 5., 7.];
v = [2., 4., 6., 8.];
w = [9., 8., 7., 6.];
X = reduce(hcat, [u, v, w])'
```

Out[9]:
```
3×4 Adjoint{Float64,Array{Float64,2}}:
 1.0  3.0  5.0  7.0
 2.0  4.0  6.0  8.0
 9.0  8.0  7.0  6.0
```

1.2 Multivariable calculus

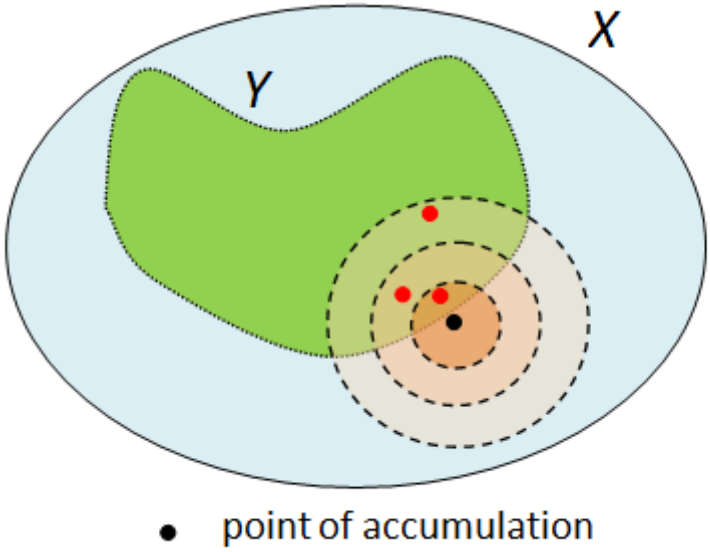
1.2.1 Limits and continuity

Throughout this section, we use the Euclidean norm $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^d x_i^2}$ for $\mathbf{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$.

The open r-ball around $\mathbf{x} \in \mathbb{R}^d$ is the set of points within Euclidean distance r of $\mathbf{x}$, that is,

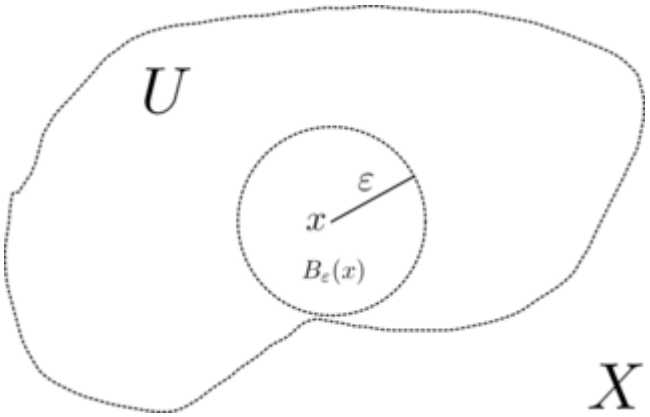
$$B_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^d \,:\, \|\mathbf{y} - \mathbf{x}\| < r\}.$$

A point $\mathbf{x} \in \mathbb{R}^d$ is a limit point (or accumulation point) of a set $A \subseteq \mathbb{R}^d$ if every open ball around $\mathbf{x}$ contains an element $\mathbf{a}$ of A such that $\mathbf{a} \neq \mathbf{x}$. A set A is closed if every limit point of A belongs to A.



(Source (https://www.math212.com/2017/12/limit-point-closure.html))

A point $\mathbf{x} \in \mathbb{R}^d$ is an interior point of a set $A \subseteq \mathbb{R}^d$ if there exists an $r > 0$ such that $B_r(\mathbf{x}) \subseteq A$. A set A is open if it consists entirely of interior points.



(Source (https://commons.wikimedia.org/wiki/File:Open_set_-_example.png))

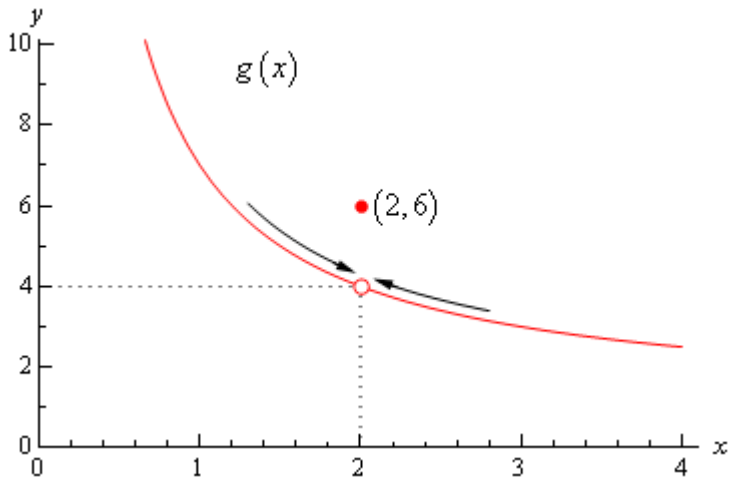
A set $A \subseteq \mathbb{R}^d$ is bounded if there exists an $r > 0$ such that $A \subseteq B_r(\mathbf{0})$, where $\mathbf{0} = (0, \ldots, 0)^T$.

Definition (Limits of a Function): Let $f : D \to \mathbb{R}$ be a real-valued function on $D \subseteq \mathbb{R}^d$. Then f is said to have a limit $L \in \mathbb{R}$ as $\mathbf{x}$ approaches $\mathbf{a}$ if: for any $\epsilon > 0$, there exists a $\delta > 0$ such that $|f(\mathbf{x}) - L| < \epsilon$ for all $\mathbf{x} \in D \cap B_\delta(\mathbf{a}) \setminus \{\mathbf{a}\}$. This is written as

$$\lim_{\mathbf{x} \to \mathbf{a}} f(\mathbf{x}) = L.$$

◁

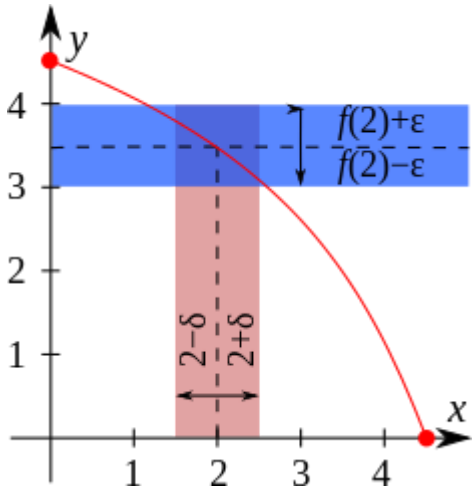
Note that we explicitly exclude $\mathbf{a}$ itself from having to satisfy the condition $|f(\mathbf{x}) - L| < \epsilon$. In particular, we may have $f(\mathbf{a}) \neq L$. We also do not restrict $\mathbf{a}$ to be in D.



Definition (Continuous Function): Let $f : D \to \mathbb{R}$ be a real-valued function on $D \subseteq \mathbb{R}^d$. Then f is said to be continuous at $\mathbf{a} \in D$ if

$$\lim_{\mathbf{x} \to \mathbf{a}} f(\mathbf{x}) = f(\mathbf{a}).$$

◁



(Source (https://commons.wikimedia.org/wiki/File:Example_of_continuous_function.svg))

We will not prove the following fundamental analysis result, which will be useful below. See e.g. Wikipedia (https://en.wikipedia.org/wiki/Extreme_value_theorem). Suppose $f : D \to \mathbb{R}$ is defined on a set $D \subseteq \mathbb{R}^d$. We say that f attains a maximum value M at $\mathbf{z}^*$ if $f(\mathbf{z}^*) = M$ and $M \geq f(\mathbf{x})$ for all $\mathbf{x} \in D$. Similarly, we say f attains a minimum value m at $\mathbf{z}_*$ if $f(\mathbf{z}_*) = m$ and $m \geq f(\mathbf{x})$ for all $\mathbf{x} \in D$.

Theorem (Extreme Value): Let $f : D \to \mathbb{R}$ be a real-valued, continuous function on a nonempty, closed, bounded set $D \subseteq \mathbb{R}^d$. Then f attains a maximum and a minimum on D.

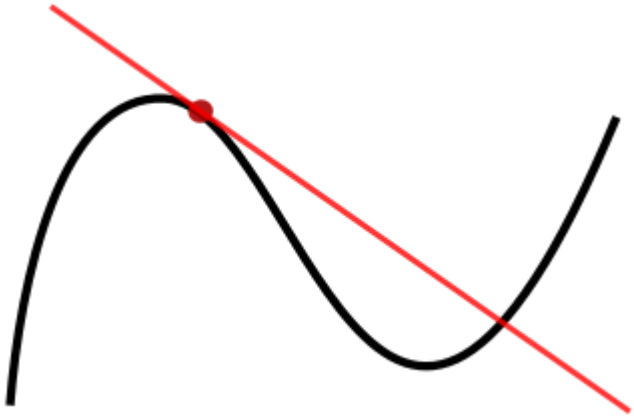
1.2.2 Derivatives

We begin by reviewing the single-variable case. Recall that the derivative of a function of a real variable is the rate of change of the function with respect to the change in the variable. Formally:

Definition (Derivative): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}$ and let $x_0 \in D$ be an interior point of D. The derivative of f at x_0 is

$$f'(x_0) = \frac{\mathrm{d}f(x_0)}{\mathrm{d}x} = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

provided the limit exists. ◁



(Source (https://commons.wikimedia.org/wiki/File:Tangent_to_a_curve.svg))

Exercise: Let f and g have derivatives at x and let α and β be constants. Show that

$$[\alpha f(x) + \beta g(x)]' = \alpha f'(x) + \beta g'(x).$$

◁

The following lemma encapsulates a key insight about the derivative of f at x_0: it tells us where to find smaller values.

Lemma (Descent Direction): Let $f : D \to \mathbb{R}$ with $D \subseteq \mathbb{R}$ and let $x_0 \in D$ be an interior point of D where $f'(x_0)$ exists. If $f'(x_0) > 0$, then there is an open ball $B_\delta(x_0) \subseteq D$ around x_0 such that for each x in $B_\delta(x_0)$:

(a) $f(x) > f(x_0)$ if $x > x_0$, (b) $f(x) < f(x_0)$ if $x < x_0$.

If instead $f'(x_0) < 0$, the opposite holds.

Proof idea: Follows from the definition of the derivative by taking ϵ small enough that $f'(x_0) - \epsilon > 0$.

Proof: Take $\epsilon = f'(x_0)/2$. By definition of the derivative, there is $\delta > 0$ such that

$$f'(x_0) - \frac{f(x_0 + h) - f(x_0)}{h} < \epsilon$$

for all $0 < h < \delta$. Rearranging gives

$$f(x_0 + h) > f(x_0) + [f'(x_0) - \epsilon]h > f(x_0)$$

by our choice of ϵ. The other direction is similar. □

For functions of several variables, we have the following generalization. As before, we let $\mathbf{e}_i \in \mathbb{R}^d$ be the i-th standard basis vector.

Definition (Partial Derivative): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^d$ and let $\mathbf{x}_0 \in D$ be an interior point of D. The partial derivative of f at $\mathbf{x}_0$ with respect to x_i is

$$\frac{\partial f(\mathbf{x}_0)}{\partial x_i} = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)}{h}$$

provided the limit exists. If $\frac{\partial f(\mathbf{x}_0)}{\partial x_i}$ exists and is continuous in an open ball around $\mathbf{x}_0$ for all i, then we say that f continuously differentiable at $\mathbf{x}_0$. ◁

Definition (Jacobian): Let $\mathbf{f} = (f_1, \ldots, f_m) : D \to \mathbb{R}^m$ where $D \subseteq \mathbb{R}^d$ and let $\mathbf{x}_0 \in D$ be an interior point of D where $\frac{\partial f_j(\mathbf{x}_0)}{\partial x_i}$ exists for all i, j. The Jacobian of $\mathbf{f}$ at $\mathbf{x}_0$ is the $d \times m$ matrix

$$\mathbf{J}_\mathbf{f}(\mathbf{x}_0) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x}_0)}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x}_0)}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x}_0)}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x}_0)}{\partial x_d} \end{pmatrix}.$$

For a real-valued function $f : D \to \mathbb{R}$, the Jacobian reduces to the row vector

$$\mathbf{J}_f(\mathbf{x}_0) = \nabla f(\mathbf{x}_0)^T$$

where the vector

$$\nabla f(\mathbf{x}_0) = \left(\frac{\partial f(\mathbf{x}_0)}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x}_0)}{\partial x_d} \right)^T$$

is the gradient of f at $\mathbf{x}_0$. ◁

Example: Consider the affine function

$$f(\mathbf{x}) = \mathbf{q}^T \mathbf{x} + r$$

where $\mathbf{x} = (x_1, \ldots, x_d)^T, \mathbf{q} = (q_1, \ldots, q_d)^T \in \mathbb{R}^d$. The partial derivatives of the linear term are given by

$$\frac{\partial}{\partial x_i}[\mathbf{q}^T \mathbf{x}] = \frac{\partial}{\partial x_i}\left[\sum_{j=1}^{d} q_j x_j\right] = q_i.$$

So the gradient of f is

$$\nabla f(\mathbf{x}) = \mathbf{q}.$$

◁

Example: Consider the quadratic function

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T P\mathbf{x} + \mathbf{q}^T \mathbf{x} + r.$$

where $\mathbf{x} = (x_1, \ldots, x_d)^T, \mathbf{q} = (q_1, \ldots, q_d)^T \in \mathbb{R}^d$ and $P \in \mathbb{R}^{d \times d}$. The partial derivatives of the quadratic term are given by

$$\begin{aligned}
\frac{\partial}{\partial x_i}[\mathbf{x}^T P\mathbf{x}] &= \frac{\partial}{\partial x_i}\left[\sum_{j,k=1}^{d} P_{jk} x_j x_k\right] \\
&= \frac{\partial}{\partial x_i}\left[P_{ii} x_i^2 + \sum_{j=1, j\neq i}^{d} P_{ji} x_j x_i + \sum_{k=1, k\neq i}^{d} P_{ik} x_i x_k\right] \\
&= 2P_{ii} x_i + \sum_{j=1, j\neq i}^{d} P_{ji} x_j + \sum_{k=1, k\neq i}^{d} P_{ik} x_k \\
&= \sum_{j=1}^{d} [P^T]_{ij} x_j + \sum_{k=1}^{d} [P]_{ik} x_k.
\end{aligned}$$

So the gradient of f is

$$\nabla f(\mathbf{x}) = \frac{1}{2}[P + P^T]\,\mathbf{x} + \mathbf{q}.$$

◁

1.2.3 Optimization

Optimization problems play an important role in data science. Here we look at unconstrained optimization problems of the form:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$$

where $f : \mathbb{R}^d \to \mathbb{R}$. Ideally, we would like to find a global minimizer to the optimization problem above.

Definition (Global Minimizer): Let $f : \mathbb{R}^d \to \mathbb{R}$. The point $\mathbf{x}^* \in \mathbb{R}^d$ is a global minimizer of f over $\mathbb{R}^d$ if

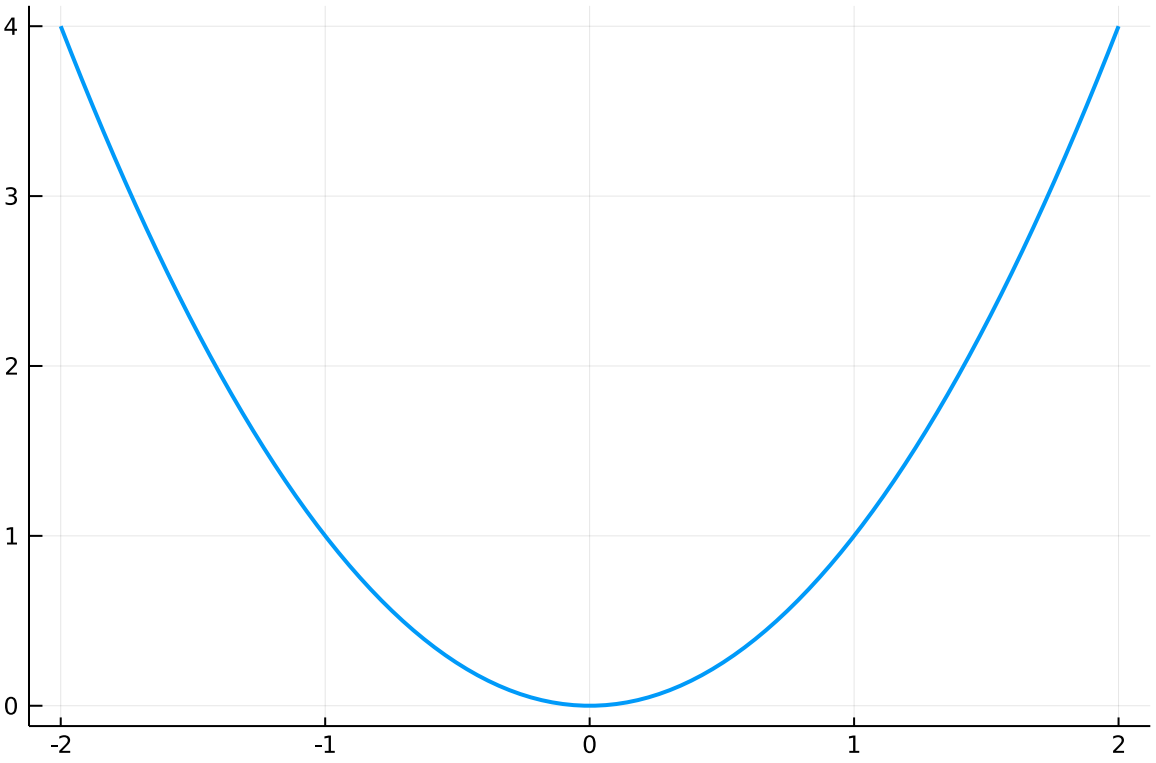
$$f(\mathbf{x}) \geq f(\mathbf{x}^*), \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

◁

Example: The function $f(x) = x^2$ over $\mathbb{R}$ has a global minimizer at $x^* = 0$. Indeed, we clearly have $f(x) \geq 0$ for all x while $f(0) = 0$.

In [10]:
```
f(x) = x^2
x = LinRange(-2,2,100)
y = f.(x)
plot(x, y, lw=2, legend=false)
```

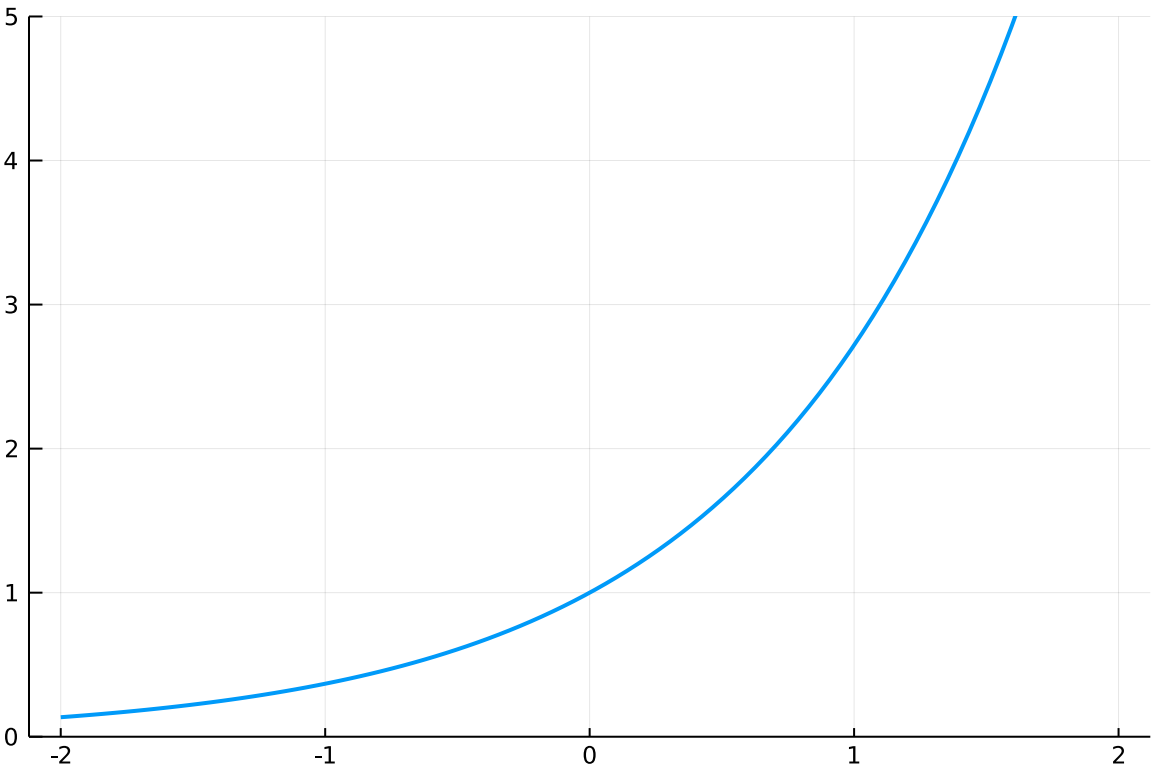
Out[10]:



The function $f(x) = e^x$ over $\mathbb{R}$ does not have a global minimizer. Indeed, $f(x) > 0$ but no x achieves 0. And, for any $m > 0$, there is x small enough such that $f(x) < m$.

In [11]:
```
f(x) = exp(x)
x = LinRange(-2,2, 100)
y = f.(x)
plot(x, y, lw=2, legend=false, ylim = (0,5))
```

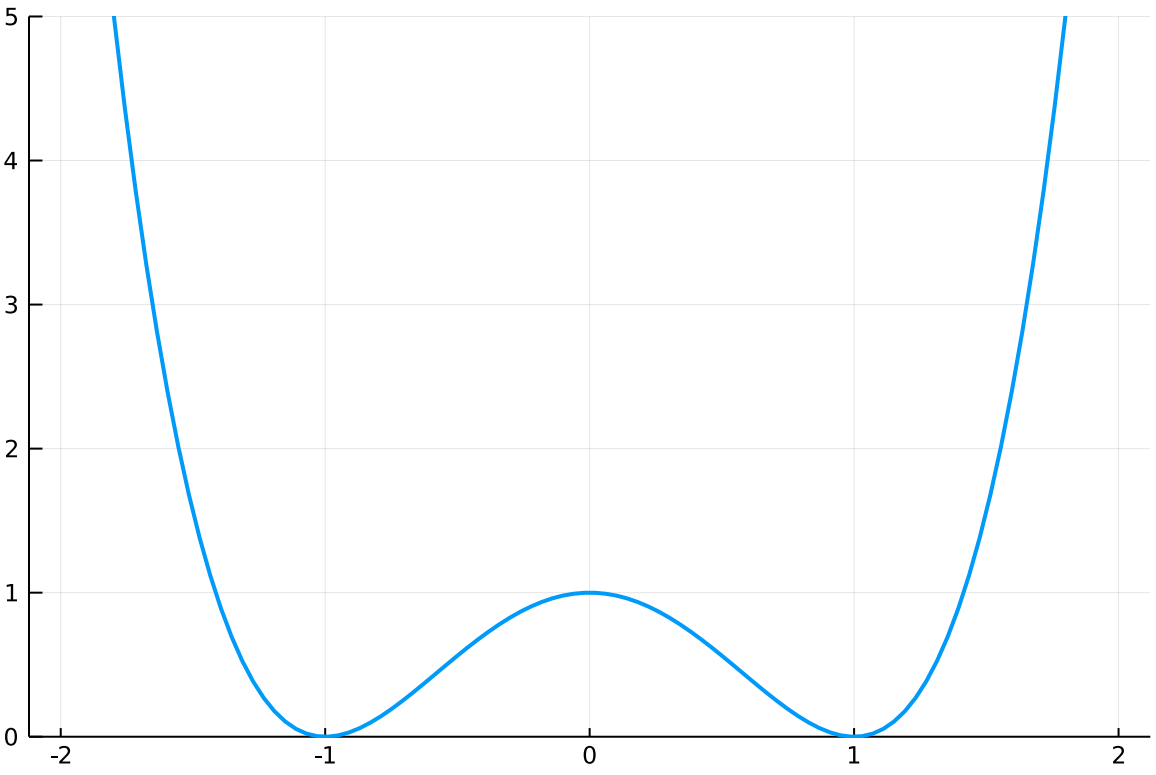
Out[11]:



The function $f(x) = (x+1)^2(x-1)^2$ over $\mathbb{R}$ has two global minimizers at $x^* = -1$ and $x^{**} = 1$. Indeed, $f(x) \geq 0$ and $f(x) = 0$ if and only $x = x^*$ or $x = x^{**}$.

In [12]:
```
f(x) = (x+1)^2*(x-1)^2
x = LinRange(-2,2, 100)
y = f.(x)
plot(x, y, lw=2, legend=false, ylim = (0,5))
```

Out[12]:



◁

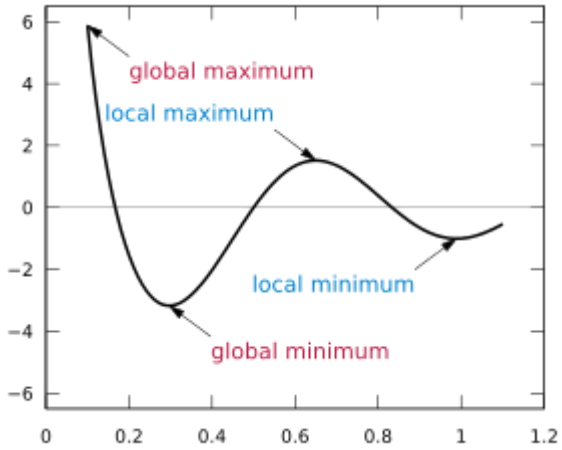
In general, finding a global minimizer and certifying that one has been found can be difficult unless some special structure is present. Therefore weaker notions of solution have been introduced.

Definition (Local Minimizer): Let $f : \mathbb{R}^d \to \mathbb{R}$. The point $\mathbf{x}^* \in \mathbb{R}^d$ is a local minimizer of f over $\mathbb{R}^d$ if there is $\delta > 0$ such that

$$f(\mathbf{x}) \geq f(\mathbf{x}^*), \quad \forall \mathbf{x} \in B_\delta(\mathbf{x}^*) \setminus \{\mathbf{x}^*\}.$$

If the inequality is strict, we say that $\mathbf{x}^*$ is a strict local minimizer. ◁

In words, $\mathbf{x}^*$ is a local minimizer if there is open ball around $\mathbf{x}^*$ where it attains the minimum value. The difference between global and local minimizers is illustrated in the next figure.



(Source (https://commons.wikimedia.org/wiki/File:Extrema_example_original.svg))

Local minimizers can be characterized in terms of the gradient, at least in terms of a necessary condition. We will prove this result later in the course.

Theorem (First-Order Necessary Condition): Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable on $\mathbb{R}^d$. If $\mathbf{x}_0$ is a local minimizer, then $\nabla f(\mathbf{x}_0) = 0$.

1.3 Probability

1.3.1 Expectation, variance and Chebyshev's inequality

Recall that the expectation (https://en.wikipedia.org/wiki/Expected_value) (or mean) of a function h of a discrete random variable X taking values in $\mathcal{X}$ is given by

$$\mathbb{E}[h(X)] = \sum_{x \in \mathcal{X}} h(x)\, p_X(x)$$

where $p_X(x) = \mathbb{P}[X = x]$ is the probability mass function (https://en.wikipedia.org/wiki/Probability_mass_function) (PMF) of X. In the continuous case, we have

$$\mathbb{E}[h(X)] = \int h(x) f_X(x)\, \mathrm{d}x$$

if f_X is the probability density function (https://en.wikipedia.org/wiki/Probability_density_function) (PDF) of X.

Two key properties of the expectation:

- *linearity*, that is,

$$\mathbb{E}[\alpha h(X) + \beta] = \alpha\, \mathbb{E}[h(X)] + \beta$$

- *monotonicity*, that is, if $h_1(x) \leq h_2(x)$ for all x then

$$\mathbb{E}[h_1(X)] \leq \mathbb{E}[h_2(X)].$$

The variance (https://en.wikipedia.org/wiki/Variance) of a real-valued random variable X is

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

and its standard deviation is $\sigma_X = \sqrt{\mathrm{Var}[X]}$. The variance does not satisfy linearity, but we have the following property

$$\mathrm{Var}[\alpha X + \beta] = \alpha^2\, \mathrm{Var}[X].$$

The variance is a measure of the typical deviation of X around its mean. A quantified version of this statement is given by Chebyshev's inequality.

Lemma (Chebyshev) For a random variable X with finite variance, we have for any $\alpha > 0$

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq \alpha] \leq \frac{\mathrm{Var}[X]}{\alpha^2}.$$

The intuition is the following: if the expected squared deviation from the mean is small, then the deviation from the mean is unlikely to be large.

To formalize this we prove a more general inequality, Makov's inequality. In words, if a non-negative random variable has a small expectation then it is unlikely to be large.

Lemma (Markov) Let Z be a non-negative random variable with finite expectation. Then, for any $\beta > 0$,

$$\mathbb{P}[Z \geq \beta] \leq \frac{\mathbb{E}[Z]}{\beta}.$$

Proof idea: The quantity $\beta\,\mathbb{P}[Z \geq \beta]$ is a lower bound on the expectation of Z restricted to the range $\{Z \geq \beta\}$, which by non-negativity is itself lower bounded by $\mathbb{E}[Z]$.

Proof: Formally, let $\mathbf{1}_A$ be the indicator of the event A, that is, it is the random variable that is 1 when A occurs and 0 otherwise. By definition, the expectation of $\mathbf{1}_A$ is

$$\mathbb{E}[A] = 0\,\mathbb{P}[\mathbf{1}_A = 0] + 1\,\mathbb{P}[\mathbf{1}_A = 1] = \mathbb{P}[A]$$

where A^c is the complement of A. Hence, by linearity and monotonicity,

$$\beta\,\mathbb{P}[Z \geq \beta] = \beta\,\mathbb{E}[\mathbf{1}_{Z \geq \beta}] = \mathbb{E}[\beta \mathbf{1}_{Z \geq \beta}] \leq \mathbb{E}[Z].$$

Rearranging gives the claim. □

Finally we return to the proof of *Chebyshev*.

Proof idea (Chebyshev): Simply apply Markov to the squared deviation of X from its mean.

Proof (Chebyshev): Let $Z = (X - \mathbb{E}[X])^2$, which is non-negative by definition. Hence, by *Markov*, for any $\beta = \alpha^2 > 0$

$$\begin{aligned}
\mathbb{P}[|X - \mathbb{E}[X]| \geq \alpha] &= \mathbb{P}[(X - \mathbb{E}[X])^2 \geq \alpha^2] \\
&= \mathbb{P}[Z \geq \beta] \\
&\leq \frac{\mathbb{E}[Z]}{\beta} \\
&= \frac{\mathrm{Var}[X]}{\alpha^2}
\end{aligned}$$

where we used the definition of the variance in the last equality.□

Chebyshev's inequality is particularly useful when combined with independence.

1.3.2 Independence and limit theorems

Recall that discrete random variables X and Y are independent if their joint PMF factorizes, that is

$$p_{X,Y}(x, y) = p_X(x)\, p_Y(y), \qquad \forall x, y$$

where $p_{X,Y}(x, y) = \mathbb{P}[X = x, Y = y]$. Similarly, continuous random variables X and Y are independent if their joint PDF factorizes. One consequence is that expectations of products of single-variable functions factorize as well, that is, for functions g and h we have

$$\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)].$$

The latter has the following important implication for the variance. If $X_1, \ldots, X_n$ are independent, real-valued random variables, then

$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n].$$

Notice that, unlike the case of the expectation, this equation for the variance requires independence in general.

Applied to the sample mean of n independent, identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$, we obtain

$$\begin{aligned}
\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[X_i] \\
&= \frac{1}{n^2} n\, \mathrm{Var}[X_1] \\
&= \frac{\mathrm{Var}[X_1]}{n}.
\end{aligned}$$

So the variance of the sample mean decreases as n gets large, while its expectation remains the same by linearity

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{n} n\, \mathbb{E}[X_1] \\
&= \mathbb{E}[X_1].
\end{aligned}$$

Together with Chebyshev's inequality, we immediately get that the sample mean approaches its expectation in the following probabilistic sense.

Theorem (Law of Large Numbers) Let $X_1, \ldots, X_n$ be i.i.d. For any $\varepsilon > 0$, as $n \to +\infty$,

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1]\right| \geq \varepsilon\right] \to 0.$$

Proof: By *Chebyshev* and the formulas above,

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X_1]\right| \geq \varepsilon\right] &= \mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right| \geq \varepsilon\right] \\
&\leq \frac{\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]}{\varepsilon^2} \\
&= \frac{\mathrm{Var}[X_1]}{n\varepsilon^2} \\
&\to 0
\end{aligned}$$

as $n \to +\infty$. □

NUMERICAL CORNER We can use simulations to confirm the *Law of Large Numbers*. Recall that a uniform random variable over the interval $[a, b]$ has density

$$f_X(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{o.w.} \end{cases}$$

We write $X \sim \mathrm{U}[a, b]$. We can obtain a sample from $\mathrm{U}[0, 1]$ by using the function `rand` (https://docs.julialang.org/en/v1.2/stdlib/Random/#Base.rand) in Julia.

In [13]:
```
rand(1)
```

Out[13]:
```
1-element Array{Float64,1}:
 0.6833180625581647
```

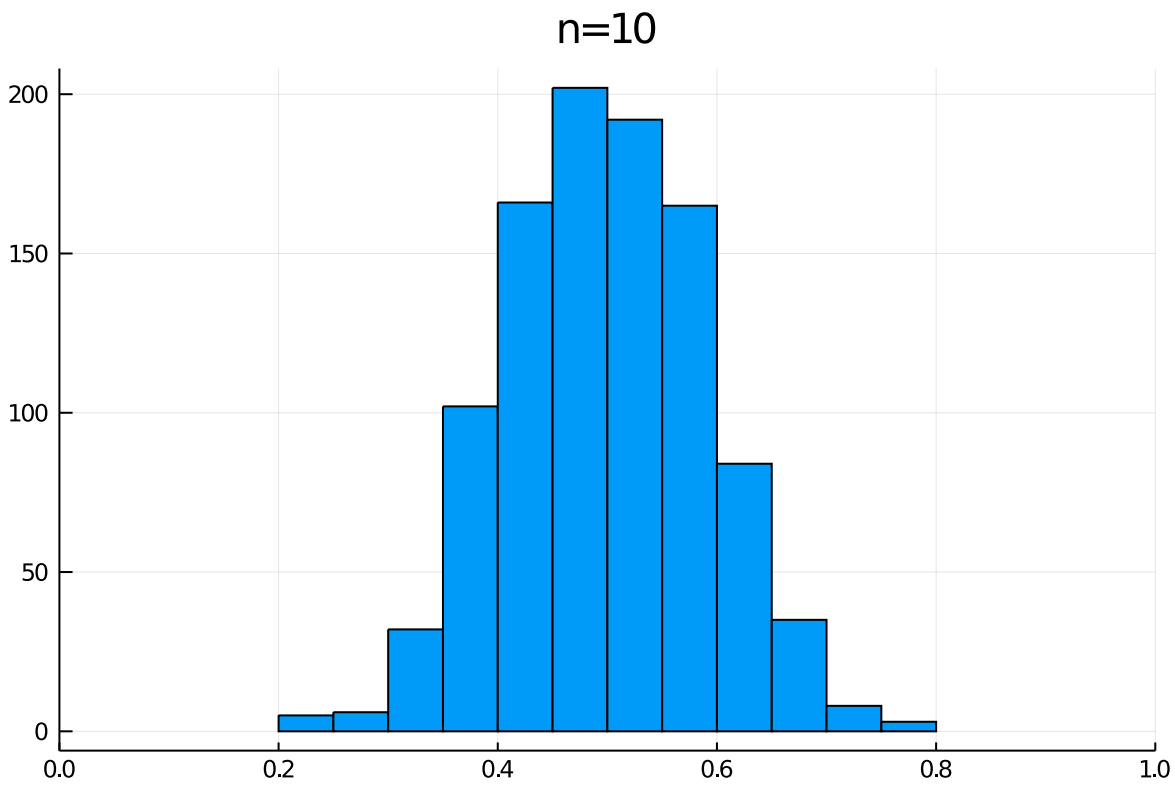
Now we take n samples from $\mathrm{U}[0, 1]$ and compute their sample mean. We repeat k times and display the empirical distribution of the sample means using an histogram (https://en.wikipedia.org/wiki/Histogram).

In [14]:
```
function lln_unif(n, k)
    sample_mean = [mean(rand(n)) for i=1:k]
    histogram(sample_mean,
        legend=false, title="n=$n", xlims=(0,1), nbin=15) # "$n" is stri
ng with value n
end
```

Out[14]:
```
lln_unif (generic function with 1 method)
```

In [15]:
```
lln_unif(10, 1000)
```

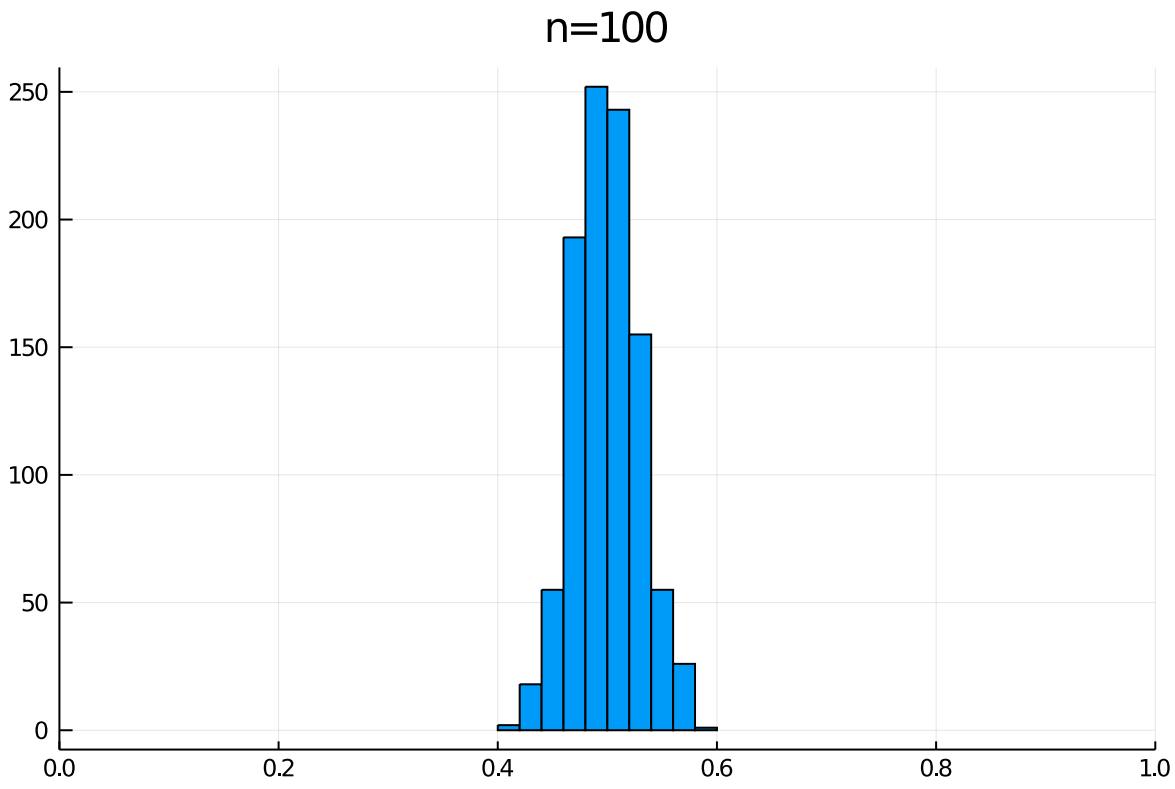
Out[15]:



Taking n much larger leads to more concentration around the mean.

In [16]:
```
lln_unif(100, 1000)
```

Out[16]:



Exercise: Recall that the cumulative distribution function (CDF) of a random variable X is defined as

$$F_X(z) = \mathbb{P}[X \leq z], \qquad \forall z \in \mathbb{R}.$$

(a) Let $\mathcal{Z}$ be the interval where $F_X(z) \in (0, 1)$ and assume that F_X is strictly increasing on $\mathcal{Z}$. Let $U \sim \mathrm{U}[0, 1]$. Show that

$$\mathbb{P}[F_X^{-1}(U) \leq z] = F_X(z).$$

(b) Generate a sample from $\mathrm{U}[a, b]$ for arbitrary a, b using `rand` and the observation in (a). This is called the inverse transform sampling method.

In [17]:
```
# Try it!
a, b = -1, 1;
X = rand(1);
# EDIT THIS LINE: transform X to obtain a random variable Y ~ U[a,b]
```

◁