TOPIC 3

Optimality, convexity, and gradient descent

4 Gradient descent

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: Oct 17, 2020

We consider a natural approach for solving optimization problems numerically: a class of algorithms known as descent methods.

Let $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable. We restrict ourselves to unconstrained minimization problems of the form

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

Ideally one would like to identify a global minimizer of f. A naive approach might be to evaluate f at a large number of points $\mathbf{x}$, say on a dense grid. However, even if we were satisfied with an approximate solution and limited ourselves to a bounded subset of the domain of f, this type of exhaustive search (https://en.wikipedia.org/wiki/Brute-force_search) is wasteful and impractical in large dimension d, as the number of points interrogated grows exponentially with d.

A less naive approach might be to find all stationary points of f, that is, those $\mathbf{x}$'s such that $\nabla f(\mathbf{x}) = \mathbf{0}$. And then choose that $\mathbf{x}$ among them that produces the smallest value of $f(\mathbf{x})$. This indeed works in many problems, like the following example we have encountered previously.

Example: Consider the least-squares problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} \|A\mathbf{x}-\mathbf{b}\|^2$$

where $A \in \mathbb{R}^{n\times d}$ has full column rank. In particular, $d \leq n$. The objective function is a quadratic function

$$f(\mathbf{x}) = \|A\mathbf{x}-\mathbf{b}\|^2 = (A\mathbf{x}-\mathbf{b})^T(A\mathbf{x}-\mathbf{b}) = \mathbf{x}^T A^T A\mathbf{x} - 2\mathbf{b}^T A\mathbf{x} + \mathbf{b}^T\mathbf{b}.$$

By a previous example,

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

where we used that $A^T A$ is symmetric.

So the stationary points satisfy

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

which you may recognize as the normal equations for the least-squares problem. We have previously shown that there is a unique solution to this system when A has full column rank. Moreover, this optimization problem is convex. Indeed, by our previous example, the Hessian of f is

$$\mathbf{H}_f(\mathbf{x}) = 2A^T A.$$

This Hessian is positive semidefinite since, for any $\mathbf{z} \in \mathbf{R}^d$,

$$\langle \mathbf{z}, 2A^T A\mathbf{z}\rangle = 2(A\mathbf{z})^T(A\mathbf{z}) = 2\|A\mathbf{z}\|^2 \geq 0.$$

So any local minimizer, which is necessarily a stationary point, is also a global minimizer. So we have found all global minimizers. ⊲

Unfortunately, identifying stationary points often leads to systems of nonlinear equations that do not have explicit solutions. Hence we resort to a different approach.

4.1 Steepest descent

In steepest descent, we attempt to find smaller and smaller values of f by successively following directions in which f decreases. As we have seen in the proof of the *First-Order Necessary Condition*, $-\nabla f$ provides such a direction. In fact, it is the direction of steepest descent in the following sense.

Recall from the *Descent Direction and Directional Derivative Lemma* that $\mathbf{v}$ is a descent direction at $\mathbf{x}_0$ if the directional derivative of f at $\mathbf{x}_0$ in the direction $\mathbf{v}$ is negative.

Lemma (Steepest Descent): Let $f : \mathbb{R}^d \to \mathbb{R}$ be continuously differentiable at $\mathbf{x}_0$. For any unit vector $\mathbf{v} \in \mathbb{R}^d$,

$$\frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{v}} \geq \frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{v}^*}$$

where

$$\mathbf{v}^* = -\frac{\nabla f(\mathbf{x}_0)}{\|\nabla f(\mathbf{x}_0)\|}.$$

Proof idea: This is an immediate application of the *Chain Rule* and *Cauchy-Schwarz*.

Proof: By the *Chain Rule* and *Cauchy-Schwarz*,

$$\begin{aligned}
\frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{v}} &= \nabla f(\mathbf{x}_0)^T \mathbf{v} \\
&\leq \|\nabla f(\mathbf{x}_0)\| \|\mathbf{v}\| \\
&= \|\nabla f(\mathbf{x}_0)\| \\
&= \nabla f(\mathbf{x}_0)^T \left(-\frac{\nabla f(\mathbf{x}_0)}{\|\nabla f(\mathbf{x}_0)\|} \right) \\
&= \frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{v}^*}.
\end{aligned}$$

□

At each interation of steepest descent, we take a step in the direction of the negative of the gradient, that is,

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \alpha_k \nabla f(\mathbf{x}^k), \quad k = 0, 1, 2 \ldots$$

for a sequence of steplengths $\alpha_k > 0$.

In general, we will not be able to guarantee that a global minimizer is reached in the limit, even if one exists. Our goal for now is more modest: to find a point where the gradient of f approximately vanishes.

NUMERICAL CORNER We implement steepest descent in Julia. We assume that a function `f` and its gradient `grad_f` are provided. We first code the basic descent step with a steplength parameter `beta`.

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

In [2]:
```
function desc_update(grad_f, x, β)
    return x .- β*grad_f(x)
end
```

Out[2]:
```
desc_update (generic function with 1 method)
```

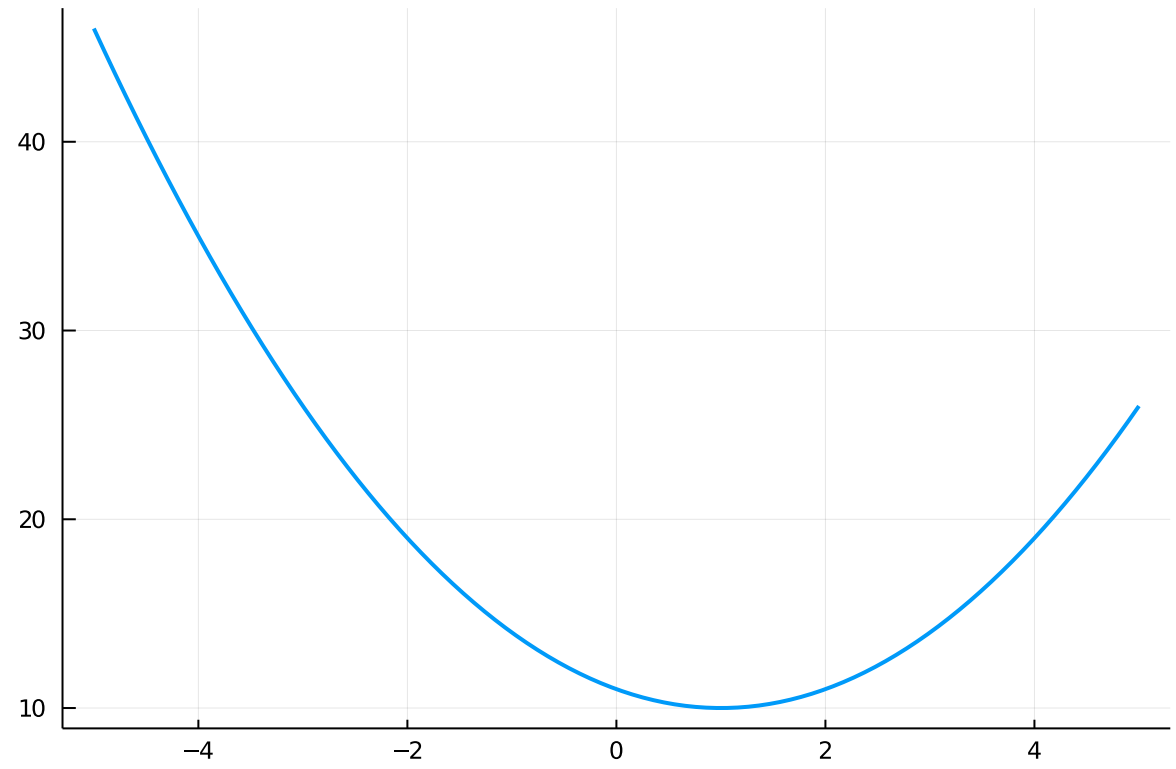
In [3]:
```
function mmids_gd(f, grad_f, x0; β=1e-3, niters=1e6)

    xk = x0 # initialization
    for _ = 1:niters
        xk = desc_update(grad_f, xk, β) # gradient step
    end

    return xk, f(xk)
end
```

Out[3]:
```
mmids_gd (generic function with 1 method)
```

We illustrate on a simple example.

In [4]:
```
f(x) = (x-1)^2+10
xgrid = LinRange(-5,5,100)
plot(xgrid, f.(xgrid), lw=2, legend=false)
```

Out[4]:



In [5]:
```
grad_f(x) = 2*(x-1)
```

Out[5]:
```
grad_f (generic function with 1 method)
```

In [6]:
```
mmids_gd(f, grad_f, 0)
```

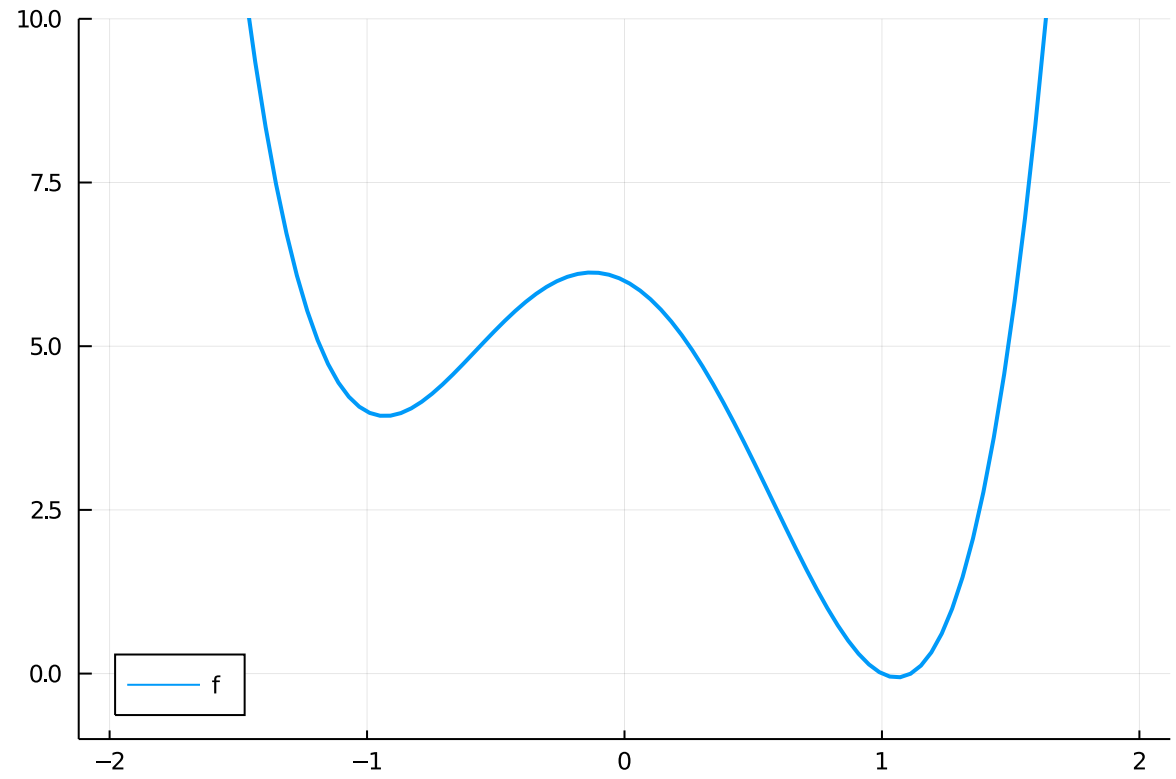
Out[6]:
```
(0.9999999999999722, 10.0)
```

We found a global minmizer in this case.

The next example shows that a different local minimizer may be reached depending on the starting point.

In [7]:
```
f(x) = 4*(x-1)^2*(x+1)^2 - 2*(x-1)
xgrid = LinRange(-2,2,100)
plot(xgrid, f.(xgrid), lw=2, label="f", ylim=(-1,10), legend=:bottomleft
)
```

Out[7]:

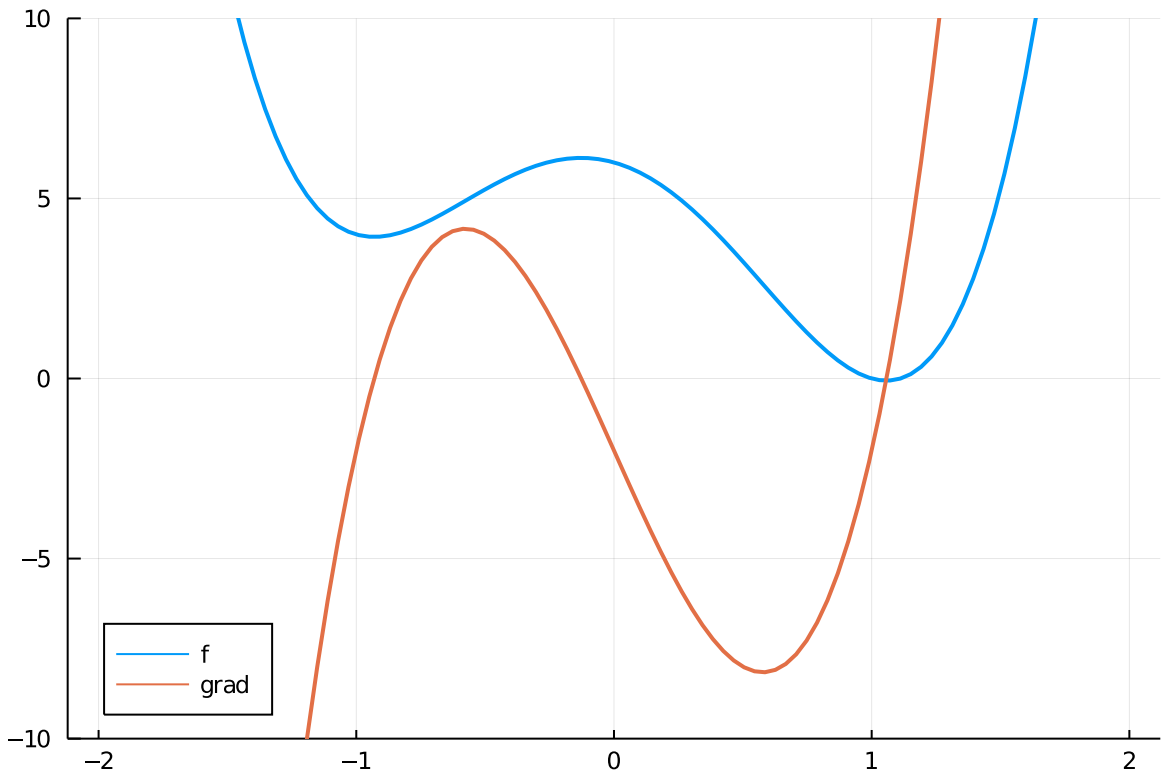


In [8]:
```
grad_f(x) = 8*(x-1)*(x+1)^2 + 8*(x-1)^2*(x+1) - 2
```

Out[8]:
```
grad_f (generic function with 1 method)
```

In [9]:
```
plot!(xgrid, grad_f.(xgrid), lw=2, label="grad", ylim=(-10,10))
```

Out[9]:



In [10]:
```
mmids_gd(f, grad_f, 0)
```

Out[10]:
```
(1.057453770738375, -0.0590145651028224)
```

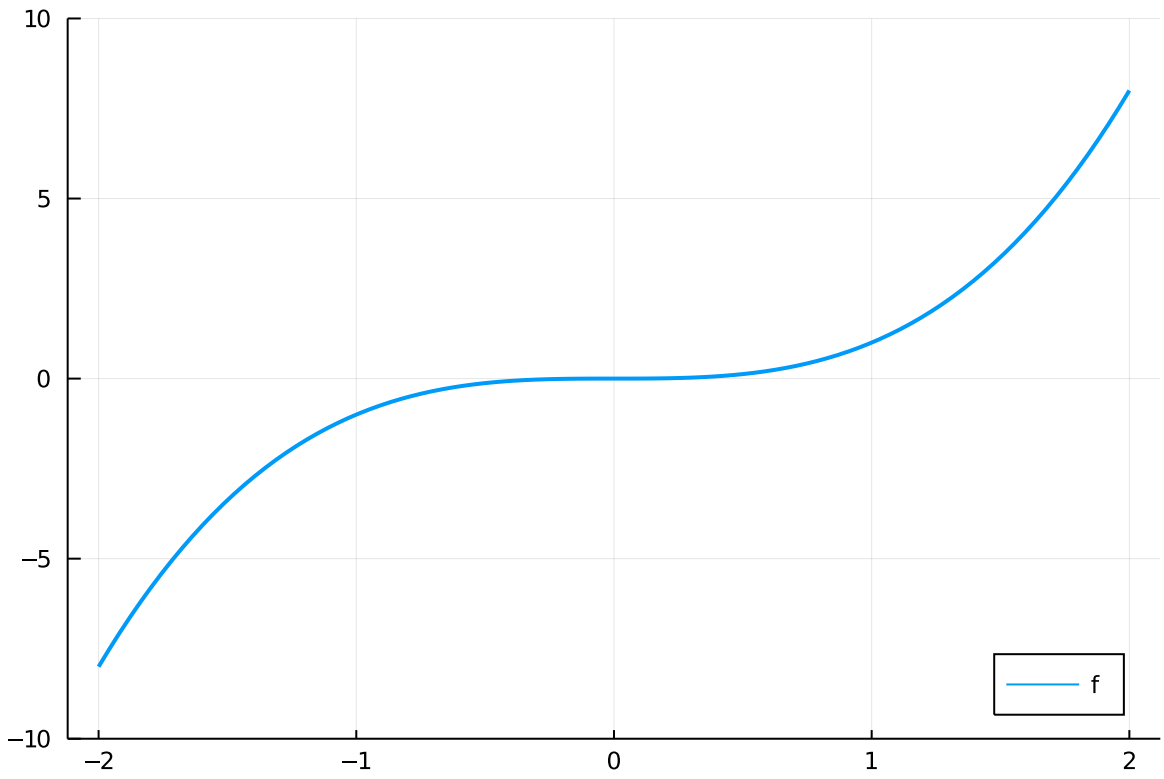
In [11]:
```
mmids_gd(f, grad_f, -2)
```

Out[11]:
```
(-0.9304029265558538, 3.933005966859003)
```

In the final example, we end up at a stationary point that is not a local minimizer. Here both the first and second derivatives are zero. This is known as a [saddle point (https://en.wikipedia.org/wiki/Saddle_point)](https://en.wikipedia.org/wiki/Saddle_point).

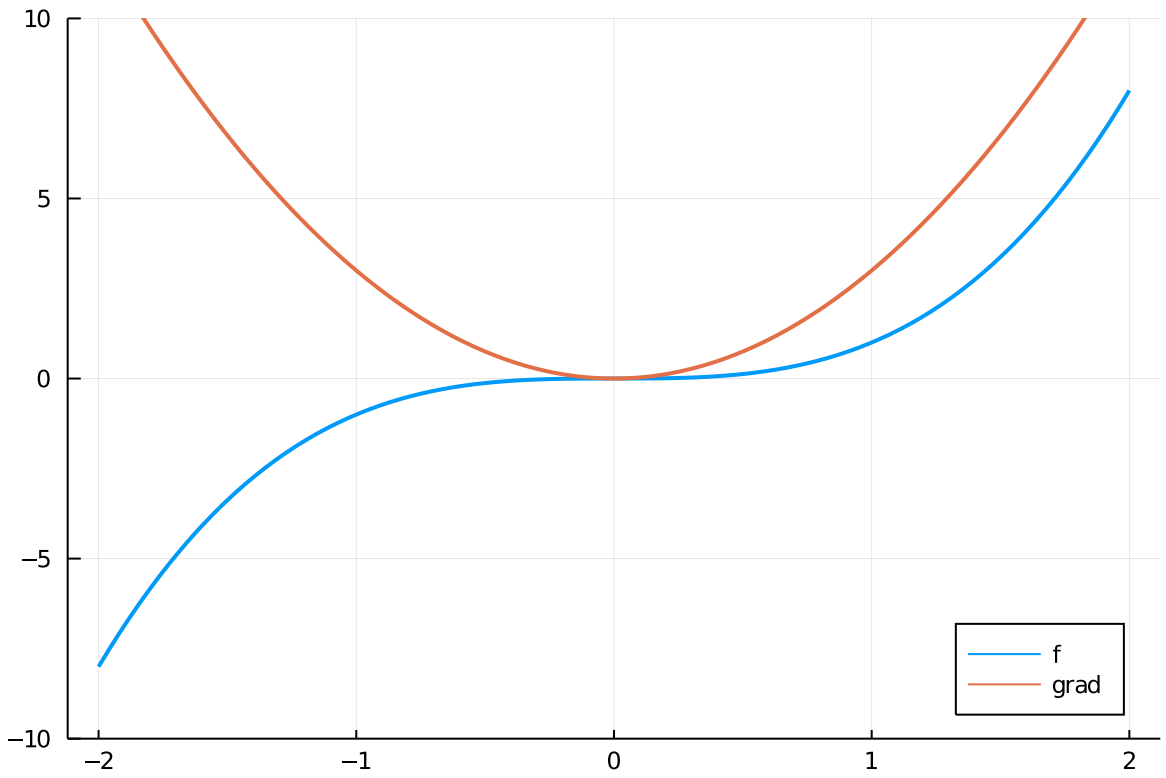
In [12]:
```
f(x) = x^3
xgrid = LinRange(-2,2,100)
plot(xgrid, f.(xgrid), lw=2, label="f", ylim=(-10,10), legend=:bottomrig
ht)
```

Out[12]:



In [13]:
```
grad_f(x) = 3*x^2
plot!(xgrid, grad_f.(xgrid), lw=2, label="grad", ylim=(-10,10))
```

Out[13]:



In [14]:
```
mmids_gd(f, grad_f, 2)
```

Out[14]:
```
(0.00033327488712690107, 3.701755838398568e-11)
```

In [15]:
```
mmids_gd(f, grad_f, -2)
```

Out[15]:
```
(-Inf, -Inf)
```

4.2 Convergence analysis

In this section, we prove some results about the convergence of steepest descent. We start with the smooth case.

4.2.1 Smooth case

We will use the following notation. Let $A, B \in \mathbb{R}^{d \times d}$ be symmetric matrices. Recall that $A \succeq 0$ means that A is PSD. We write $A \preceq B$ (respectively $A \succeq B$) to indicate that $B - A \succeq 0$ (respectively $A - B \succeq 0$). We will need the following statement, whose proof is left as an exercise.

Exercise: Let $A \in \mathbb{R}^{d\times d}$ be a symmetric matrix. Show that $A \preceq M I_{d\times d}$ if and only if the eigenvalues of A are at most M. Similarly, $m I_{d\times d} \preceq A$ if and only if the eigenvalues of A are at least m. [Hint: Observe that the eigenvectors of A are also eigenvectors of the identity matrix $I_{d\times d}$.] ◁

Definition (Smooth Function): Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. We say that f is L-smooth if

$$-L I_{d\times d} \preceq \mathbf{H}_f(\mathbf{x}) \preceq L I_{d\times d}, \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

◁

Lemma (Quadratic Bound for Smooth Functions): Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. Then f is L-smooth if and only if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ it holds that

$$\left| f(\mathbf{y}) - \{ f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}) \} \right| \leq \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|^2.$$

Proof idea: We apply the *Multivariate Taylor's Theorem*, then use the extremal characterization of the eigenvalues to bound the second-order term.

Proof: By the *Multivariate Taylor's Theorem*, for any $\alpha > 0$ there is $\xi_\alpha \in (0, 1)$

$$f(\mathbf{x} + \alpha \mathbf{p}) = f(\mathbf{x}) + \alpha \nabla f(\mathbf{x})^T \mathbf{p} + \frac{1}{2} \alpha^2 \mathbf{p}^T \, \mathbf{H}_f(\mathbf{x} + \xi_\alpha \alpha \mathbf{p}) \, \mathbf{p}$$

where $\mathbf{p} = \mathbf{y} - \mathbf{x}$.

If f is L-smooth, then at $\alpha = 1$ by *Courant-Fischer*

$$-L \|\mathbf{p}\|^2 \leq \mathbf{p}^T \, \mathbf{H}_f(\mathbf{x} + \xi_1 \mathbf{p}) \, \mathbf{p} \leq L \|\mathbf{p}\|^2.$$

That implies the inequality in the statement.

On the other hand, if that inequality holds, by combining with the Taylor expansion above we get

$$\left| \frac{1}{2} \alpha^2 \mathbf{p}^T \, \mathbf{H}_f(\mathbf{x} + \xi_\alpha \alpha \mathbf{p}) \, \mathbf{p} \right| \leq \frac{L}{2} \alpha^2 \|\mathbf{p}\|^2$$

where we used that $\|\alpha \mathbf{p}\| = \alpha \|\mathbf{p}\|$ by homogeneity of the norm. Dividing by $\alpha^2/2$, then taking $\alpha \to 0$ and using the continuity of the Hessian gives

$$\left| \mathbf{p}^T \, \mathbf{H}_f(\mathbf{x}) \, \mathbf{p} \right| \leq L \|\mathbf{p}\|^2.$$

By *Courant-Fischer* again, that implies that f is L-smooth. □

We show next that, in the smooth case, steepest descent with an appropriately chosen steplength produces a sequence of points whose objective values decrease and whose gradients vanish in the limit. We also give a quantitative convergence rate.

Theorem (Convergence of Steepest Descent in Smooth Case): Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and bounded from below, that is, there $\bar{f} > -\infty$ such that $f(\mathbf{x}) \geq \bar{f}$, $\forall \mathbf{x} \in \mathbb{R}^d$. Then steepest descent with $\alpha = 1/L$ started from any $\mathbf{x}^0$ produces a sequence $\mathbf{x}^t$, $t = 1, 2, \ldots$ such that

$$f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t), \quad \forall t$$

and

$$\lim_{t \to +\infty} \|\nabla f(\mathbf{x}^t)\| = 0.$$

Moreover, after H steps, there is a t in $\{0, \ldots, H\}$ such that

$$\|\nabla f(\mathbf{x}^t)\| \leq \sqrt{\frac{2L\left[f(\mathbf{x}^0) - \bar{f}\right]}{H}}.$$

The heart of the proof is the following fundamental inequality. It also explains the choice of steplength.

Lemma (Descent Guarantee in the Smooth Case): Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth. For any $\mathbf{x} \in \mathbb{R}^d$,

$$f(\mathbf{x} - (1/L)\nabla f(\mathbf{x})) \leq f(\mathbf{x}) - \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2.$$

Proof idea (Descent Guarantee in the Smooth Case): Intuitively, the *Quadratic Bound for Smooth Functions* shows that f is well approximated by a quadratic function in a neighborhood of $\mathbf{x}$ whose size depends on the smoothness parameter L. Choosing a step that minimizes this approximation leads to a guaranteed improvement.

Proof (Descent Guarantee in the Smooth Case): By the *Quadratic Bound for Smooth Functions*, letting $\mathbf{p} = -\nabla f(\mathbf{x})$

$$\begin{aligned} f(\mathbf{x} + \alpha\mathbf{p}) &\leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\alpha\mathbf{p}) + \frac{L}{2}\|\alpha\mathbf{p}\|^2 \\ &= f(\mathbf{x}) - \alpha\|\nabla f(\mathbf{x})\|^2 + \alpha^2\frac{L}{2}\|\nabla f(\mathbf{x})\|^2 \\ &= f(\mathbf{x}) + \left(-\alpha + \alpha^2\frac{L}{2}\right)\|\nabla f(\mathbf{x})\|^2. \end{aligned}$$

The quadratic function in parentheses is convex and minimized at the stationary point α satisfying

$$\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(-\alpha + \alpha^2\frac{L}{2}\right) = -1 + \alpha L = 0.$$

Taking $\alpha = 1/L$ and replacing in the inequality above gives

$$f(\mathbf{x} - (1/L)\nabla f(\mathbf{x})) \leq f(\mathbf{x}) - \frac{1}{2L}\|\nabla f(\mathbf{x})\|^2$$

as claimed. □

We give a numerical example below using a special case of logistic regression. But first a calculus exercise.

Exercise: For $a \in [-1, 1]$ and $b \in \{0, 1\}$, let $\hat{f}(x, a) = \sigma(xa)$ where

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

be a classifier parametrized by $x \in \mathbb{R}$. For a dataset $a_i \in [-1, 1]$ and $b_i \in \{0, 1\}$, $i = 1, \ldots, n$, let the cross-entropy loss be

$$\mathcal{L}(x, \{(a_i, b_i)\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n \ell(x, a_i, b_i)$$

where

$$\ell(x, a, b) = -b\log(\hat{f}(x, a)) - (1 - b)\log(1 - \hat{f}(x, a)).$$

(a) Show that $\sigma'(t) = \sigma(t)(1 - \sigma(t))$ for all $t \in \mathbb{R}$.

(b) Use (a) to show that

$$\frac{\partial}{\partial x}\mathcal{L}(x, \{(a_i, b_i)\}_{i=1}^n) = -\frac{1}{n}\sum_{i=1}^n (b_i - \hat{f}(x, a_i))\, a_i.$$

(c) Use (b) to show that

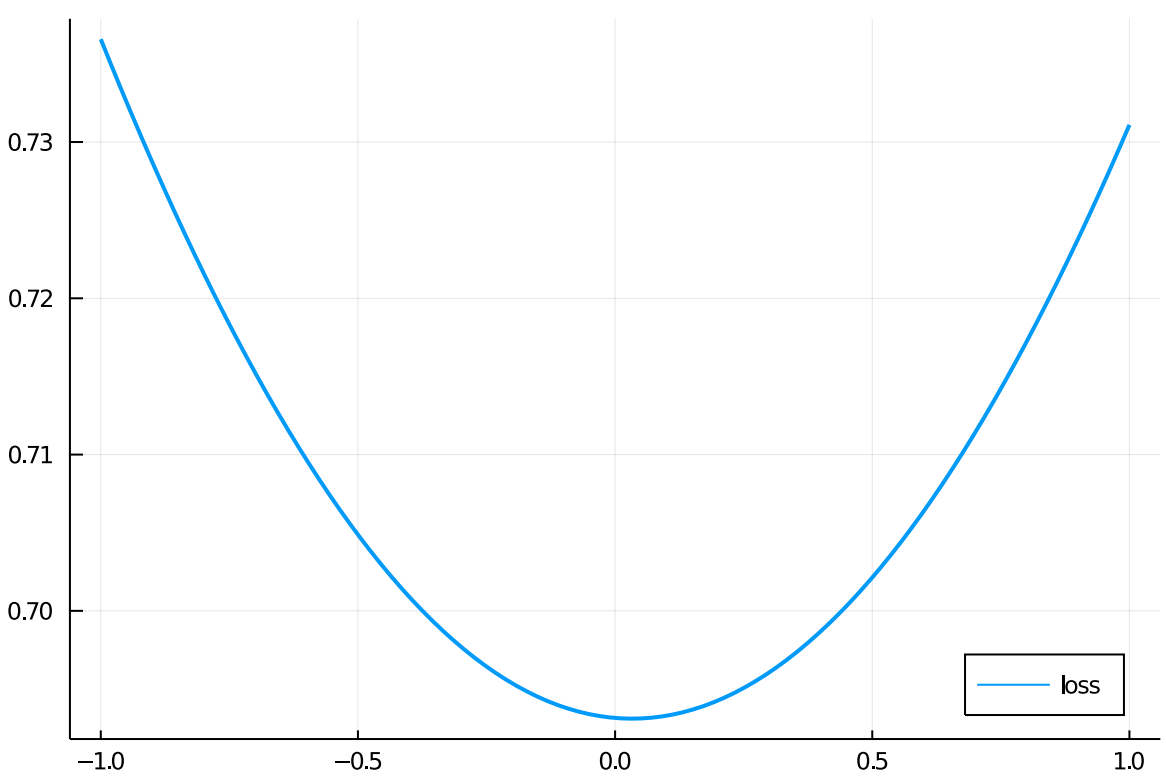
$$\frac{\partial^2}{\partial x^2}\mathcal{L}(x, \{(a_i, b_i)\}_{i=1}^n) = \frac{1}{n}\sum_{i=1}^n \hat{f}(x, a_i)\,(1 - \hat{f}(x, a_i))\, a_i^2.$$

(d) Use (c) to show that, for any dataset $\{(a_i, b_i)\}_{i=1}^n$, $\mathcal{L}$ is 1-smooth as a function of x. ◁

NUMERICAL CORNER We illustrate numerically the exercise above on a random dataset. The functions $\hat{f}$, $\mathcal{L}$ and $\frac{\partial}{\partial x}\mathcal{L}$ are defined next.

In [16]:
```
n = 10000
a = 2*rand(n) .- 1
b = rand(0:1,n)
fhat(x) = 1 ./ ( 1 .+ exp.(-x.*a))
loss(x) = mean(-b.*log.(fhat(x)) - (1 .- b).*log.(1 .- fhat(x)))
grad(x) = -mean((b .- fhat(x)).*a)
x = LinRange(-1,1,100)
plot(x, loss.(x), lw=2, label="loss", legend=:bottomright)
```

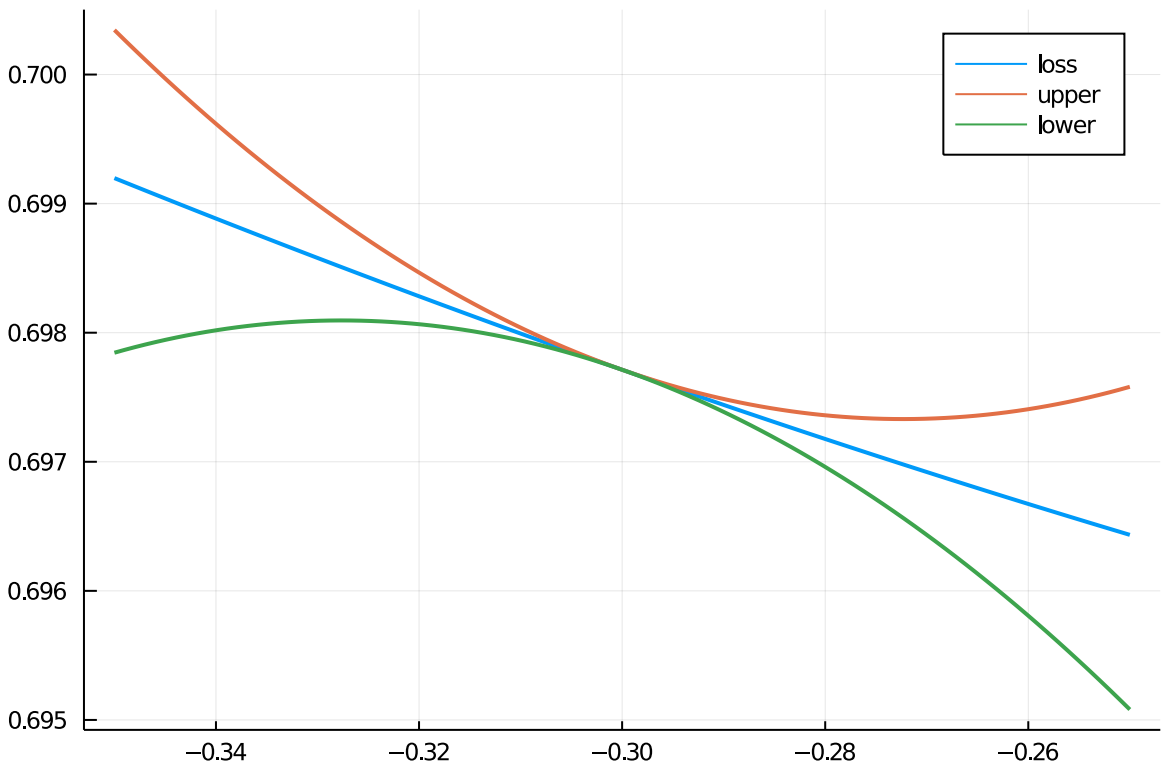
Out[16]:



We plot next the upper and lower bounds in the *Quadratic Bound for Smooth Functions* around $x = x_0$. By the previous exercise, we can take $L = 1$. Observe that minimizing the upper quadratic bound leads to a decrease in $\mathcal{L}$.

In [17]:
```
x0 = -0.3
x = LinRange(x0-0.05,x0+0.05,100)
upper = loss(x0) .+ (x .- x0)*grad(x0) .+ (1/2)*(x .- x0).^2 # upper app
roximation
lower = loss(x0) .+ (x .- x0)*grad(x0) .- (1/2)*(x .- x0).^2 # lower app
roximation
plot(x, loss.(x), lw=2, label="loss")
plot!(x, upper, lw=2, label="upper")
plot!(x, lower, lw=2, label="lower")
```

Out[17]:



◁

We return to the proof of the theorem.

Proof idea (Convergence of Steepest Descent in Smooth Case): We use a telescoping argument to write $f(\mathbf{x}^H)$ as a sum of stepwise increments, each of which can be bounded by the previous lemma. Because $f(\mathbf{x}^H)$ is bounded from below, it then follows that the gradients must vanish in the limit.

Proof (Convergence of Steepest Descent in Smooth Case): By the *Descent Guarantee in the Smooth Case*,

$$f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) - \frac{1}{2L}\|\nabla f(\mathbf{x}^t)\|^2 \leq f(\mathbf{x}^t), \quad \forall t.$$

Furthermore, using a telescoping sum, we get

$$\begin{aligned} f(\mathbf{x}^H) &= f(\mathbf{x}^0) + \sum_{t=0}^{H-1} [f(\mathbf{x}^{t+1}) - f(\mathbf{x}^t)] \\ &= f(\mathbf{x}^0) - \frac{1}{2L} \sum_{t=0}^{H-1} \|\nabla f(\mathbf{x}^t)\|^2. \end{aligned}$$

Rearranging and using $f(\mathbf{x}^H) \geq \bar{f}$ leads to

$$\sum_{t=0}^{H-1} \|\nabla f(\mathbf{x}^t)\|^2 \leq 2L[f(\mathbf{x}^0) - \bar{f}].$$

So as $H \to +\infty$, we must have $\|\nabla f(\mathbf{x}^H)\|^2 \to 0$. We also get the more quantitative bound

$$\min_{t=0,\ldots,H-1} \|\nabla f(\mathbf{x}^t)\|^2 \leq \frac{2L[f(\mathbf{x}^0) - \bar{f}]}{H}$$

as the minimum is necessarily less or equal than the average. That proves the last claim. □

4.2.2 Smooth and strongly convex case

With stronger assumptions, we can obtained stronger convergence results. One such assumption is strong convexity, which we define next. Compare the definition with that of a smooth function.

Definition (Strongly Convex Function): Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and let $m > 0$. We say that f is m-strongly convex if

$$\mathbf{H}_f(\mathbf{x}) \succeq mI_{d\times d}, \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

◁

Exercise: Consider the quadratic function

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

where P is symmetric. Show that, if P is positive definite, then f is strongly convex. ◁

The proof of the following lemma is left as an exercise.

Lemma (Quadratic Bound for Strongly Convex Functions): Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable. Then f is m-strongly convex if and only if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ it holds that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Exercise: Prove the *Quadratic Bound for Strongly Convex Functions*. [Hint: Adapt the proof of the *Quadratic Bound for Smooth Functions*.] ◁

The previous lemma immediately leads to the following fundamental result.

Theorem (Global Minimizer of Strongly Convex Functions): Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable and m-strongly convex with $m > 0$. If $\nabla f(\mathbf{x}^*) = \mathbf{0}$, then $\mathbf{x}^*$ is a unique global minimizer of f.

Proof: If $\nabla f(\mathbf{x}^*) = \mathbf{0}$, by the *Quadratic Bound for Strongly Convex Functions*,

$$f(\mathbf{y}) \geq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T(\mathbf{y} - \mathbf{x}^*) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}^*\|^2 > f(\mathbf{x}^*)$$

for all $\mathbf{y} \neq \mathbf{x}^*$, which proves the claim. □

We are now ready to prove our convergence result for smooth, strongly convex functions. We will show something stronger this time. We will control the value of f itself and obtain a much faster rate of convergence.

If f is m-strongly convex and has a global minimizer $\mathbf{x}^*$, then the global minimizer is unique and characterized by $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Strong convexity allows us to relate the value of the function at a point $\mathbf{x}$ and the gradient of f at that point. This is proved in the following lemma, which is key to our convergence results.

Lemma (Relating f and ∇f): Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice continuously differentiable, m-strongly convex with a global minimizer at $\mathbf{x}^*$. Then for any $\mathbf{x} \in \mathbb{R}^d$

$$f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{\|\nabla f(\mathbf{x})\|^2}{2m}.$$

Proof: By the *Quadratic Bound for Strongly Convex Functions*,

$$\begin{aligned} f(\mathbf{x}^*) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{x}^* - \mathbf{x}) + \frac{m}{2}\|\mathbf{x}^* - \mathbf{x}\|^2 \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T\mathbf{w} + \frac{1}{2}\mathbf{w}^T(mI_{d\times d})\mathbf{w} \end{aligned}$$

where on the second line we defined $\mathbf{w} = \mathbf{x}^* - \mathbf{x}$. The right-hand side is a quadratic function in $\mathbf{w}$ (for $\mathbf{x}$ fixed). So the inequality is still valid if we replace $\mathbf{w}$ with the global minimizer $\mathbf{w}^*$ of that quadratic function.

The matrix $mI_{d\times d}$ is positive definite since its eigenvalues are all equal to $m > 0$. By a previous example, we know that

$$\mathbf{w}^* = -(mI_{d\times d})^{-1}\nabla f(\mathbf{x}) = -\frac{1}{m}\nabla f(\mathbf{x}).$$

So, replacing $\mathbf{w}$ with $\mathbf{w}^*$, we have the inequality

$$\begin{aligned} f(\mathbf{x}^*) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T\left\{-\frac{1}{m}\nabla f(\mathbf{x})\right\} + \frac{1}{2}\left\{-\frac{1}{m}\nabla f(\mathbf{x})\right\}^T(mI_{d\times d})\left\{-\frac{1}{m}\nabla f(\mathbf{x})\right\} \\ &= f(\mathbf{x}) - \frac{1}{2m}\|\nabla f(\mathbf{x})\|^2. \end{aligned}$$

Rearranging gives the claim. □

We can now state our convergence result.

Theorem (Convergence of Steepest Descent in Smooth Case): Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and m-strongly convex with a global minimizer at $\mathbf{x}^*$. Then steepest descent with $\alpha = 1/L$ started from any $\mathbf{x}^0$ produces a sequence $\mathbf{x}^t$, $t = 1, 2, \ldots$ such that

$$\lim_{t\to+\infty} f(\mathbf{x}^t) = f(\mathbf{x}^*).$$

Moreover, after H steps, we have

$$f(\mathbf{x}^H) - f(\mathbf{x}^*) \leq \left(1 - \frac{m}{L}\right)^H [f(\mathbf{x}^0) - f(\mathbf{x}^*)].$$

Observe that $f(\mathbf{x}^H) - f(\mathbf{x}^*)$ decreases exponentially fast in H.

Proof idea (Convergence of Steepest Descent in Smooth Case): We apply the *Descent Guarantee for Smooth Functions* together with the lemma above.

Proof (Convergence of Steepest Descent in Smooth Case): By the *Descent Guarantee for Smooth Functions* together and the lemma above, we have for all t

$$f(\mathbf{x}^{t+1}) \leq f(\mathbf{x}^t) - \frac{1}{2L}\|\nabla f(\mathbf{x}^t)\|^2 \leq f(\mathbf{x}^t) - \frac{m}{L}[f(\mathbf{x}^t) - f(\mathbf{x}^*)].$$

Subtracting $f(\mathbf{x}^*)$ on both sides gives

$$f(\mathbf{x}^{t+1}) - f(\mathbf{x}^*) \leq \left(1 - \frac{m}{L}\right)[f(\mathbf{x}^t) - f(\mathbf{x}^*)].$$

Recursing gives the claim. □