Unconstrained Optimization

ME598/494 Lecture

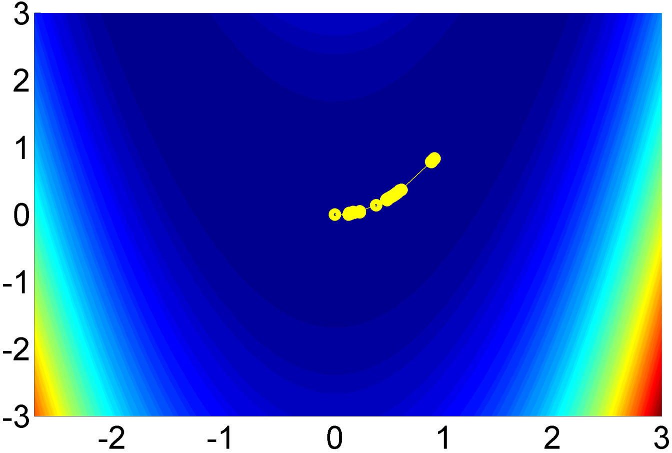
Max Yi Ren

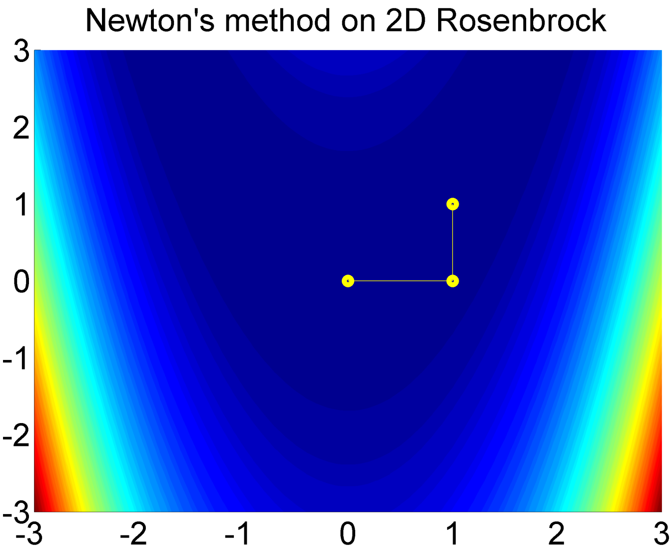
Department of Mechanical Engineering, Arizona State University

September 5, 2017

- You feel the slope
- You have a flashlight
- Which direction will you go?
- How much will you go in that direction?

Gradient descent on 2D Rosenbrock function




Newton's method on 2D Rosenbrock
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

Taylor series

Assuming function $f(x)$ has derivatives of any order, the Taylor series expansion of x about the point x_0 is

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n,$$

where $f^{(n)}(x_0)$ is the nth-order derivative at x_0.

Taylor's theorem

Let $N \geq 1$ be an integer and let the function $f(x)$ be N times differentiable at the point x_0, then

$$f(x) = f(x_0) + \sum_{n=1}^{N} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + o\left(|x - x_0|^N\right),$$

The notation for the remainder, $o\left(|x - x_0|^N\right)$, means that the remainder is small compared to $|x - x_0|^N$. Formally, $\lim_{x \to x_0} \frac{o\left(|x-x_0|^N\right)}{|x-x_0|^N} = 0$.

Local approximation

Approximations in $\mathbb{R}$:

$$(linear) \quad f(x) \approx f(x_0) + \frac{df(x_0)}{dx}(x - x_0);$$

$$(quadratic) \quad f(x) \approx f(x_0) + \frac{df(x_0)}{dx}(x - x_0) + \frac{1}{2}\frac{d^2f(x_0)}{dx^2}(x - x_0)^2$$

Approximations in $\mathbb{R}^n$

$$(linear) \quad f(\mathbf{x}) \approx f(\mathbf{x}_0) + \sum_{i=1}^{n} \frac{\partial f(\mathbf{x}_0)}{\partial x_i}(x_i - x_{i0});$$

$$(quadratic) \quad f(\mathbf{x}) \approx f(\mathbf{x}_0) + \sum_{i=1}^{n} \frac{\partial f(\mathbf{x}_0)}{\partial x_i}(x_i - x_{i0}) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j}(x_i - x_{i0})(x_j - x_{j0})$$

Local approximation

The vector form of the approximations:

linear:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \mathbf{g}_{\mathbf{x}_0}^T(\mathbf{x} - \mathbf{x}_0);$$

and quadratic:

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \mathbf{g}_{\mathbf{x}_0}^T(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T\mathbf{H}_{\mathbf{x}_0}(\mathbf{x} - \mathbf{x}_0).$$

Here $\mathbf{g}_{\mathbf{x}_0}$ and $\mathbf{H}_{\mathbf{x}_0}$ are the *gradient* and *Hessian matrix* of $f(\mathbf{x})$ at $\mathbf{x}_0$. $\mathbf{H}$ is *symmetric*.

We call $\partial f = f(\mathbf{x}) - f(\mathbf{x}_0)$ and $\partial\mathbf{x} = \mathbf{x} - \mathbf{x}_0$ the *function perturbation* and *perturbation vector* (at $\mathbf{x}_0$).

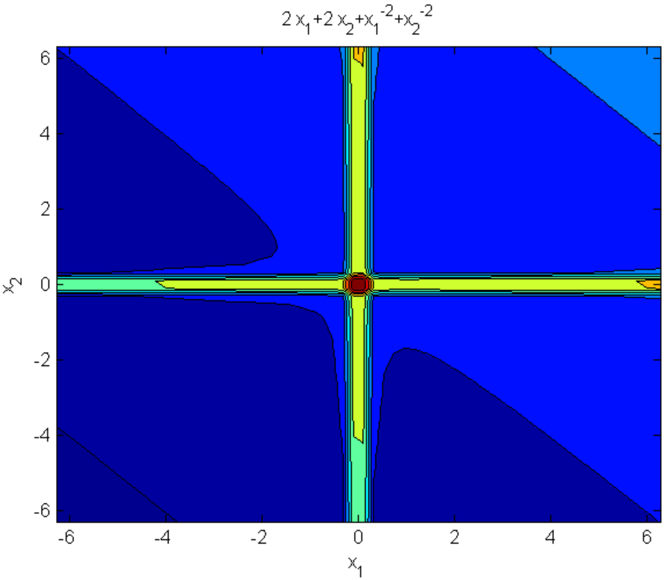
Local approximation

Exercise 1: Find the second-order "approximation" for $f(\mathbf{x}) = (3 - x_1)^2 + (4 - x_2)^2$. How many local minima do we have? What is special about the Hessian?

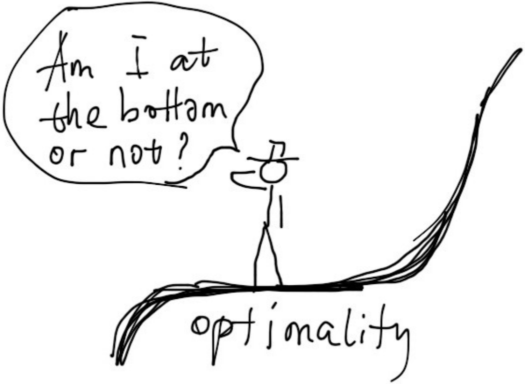
Exercise 2: Find the quadratic approximation of the function:

$$f(\mathbf{x}) = 2x_1 + x_1^{-2} + 2x_2 + x_2^{-2}, \quad \mathbf{x} \in \mathbb{R}^2, \mathbf{x} \neq (0, 0)^T.$$

Is the Hessian positive definite? Is the problem bounded?



The function has positive definite Hessian everywhere in its feasible domain, but its function value is unbounded.


Am I at
the bottom
or not?
optimality

Necessary and sufficient conditions

first-order necessary condition

If $f(\mathbf{x})$, $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n$, has a local minimum at an interior point $\mathbf{x}_*$ of the set $\mathcal{X}$ and if $f(\mathbf{x})$ is continuously differentiable at $\mathbf{x}_*$, then $\mathbf{g}_{\mathbf{x}_*} = \mathbf{0}$.

second-order optimality condition

Let $f(\mathbf{x})$ be twice differentiable at the point $\mathbf{x}_*$.

1. *(necessity) If $\mathbf{x}_*$ is a local solution, then $\mathbf{g}_{\mathbf{x}_*} = \mathbf{0}$ and Hessian at $\mathbf{x}_*$ is positive-semidefinite.*
2. *(sufficiency) If the Hessian of $f(x)$ is positive-definite at a stationary point $\mathbf{x}_*$, i.e., $\mathbf{g}_{\mathbf{x}_*} = \mathbf{0}$, then $\mathbf{x}_*$ is a local minimum.*

Exercise 3: Find the solution(s) for
$\min_{\mathbf{x}\in\mathbb{R}^2} \quad f(\mathbf{x}) = 4x_1^2 - 4x_1x_2 + x_2^2 - 4x_1 + 2x_2$.

What about first-order sufficient condition?

Proof for first-order necessary condition

From first-order approximation at $\mathbf{x}_*$ we have:

$$f(\mathbf{x}) = f(\mathbf{x}_*) + \mathbf{g}_{\mathbf{x}_*}^T(\mathbf{x} - \mathbf{x}_*) + o(||\mathbf{x} - \mathbf{x}_*||). \tag{1}$$

Let $\mathbf{x} = \mathbf{x}_* - t\mathbf{g}_{\mathbf{x}_*}$. (Here we deliberately pick $-\mathbf{g}_{\mathbf{x}_*}$ as the direction.) Since $\mathbf{x}_*$ is a local solution, we have $f(\mathbf{x}_* - t\mathbf{g}_{\mathbf{x}_*}) - f(\mathbf{x}_*) \geq 0,\ \forall t > 0$. Take Equation (1) into account to have:

$$0 \leq \frac{f(\mathbf{x}_* - t\mathbf{g}_{\mathbf{x}_*}) - f(\mathbf{x}_*)}{t} = -||\mathbf{g}_{\mathbf{x}_*}||^2 + \frac{o(t||\mathbf{g}_{\mathbf{x}_*}||)}{t}.$$

Taking $t \to 0$, we have $0 \leq -||\mathbf{g}_{\mathbf{x}_*}||^2 \leq 0$, requiring $\mathbf{g}_{\mathbf{x}_*} = 0$.

Proof for second-order necessary condition

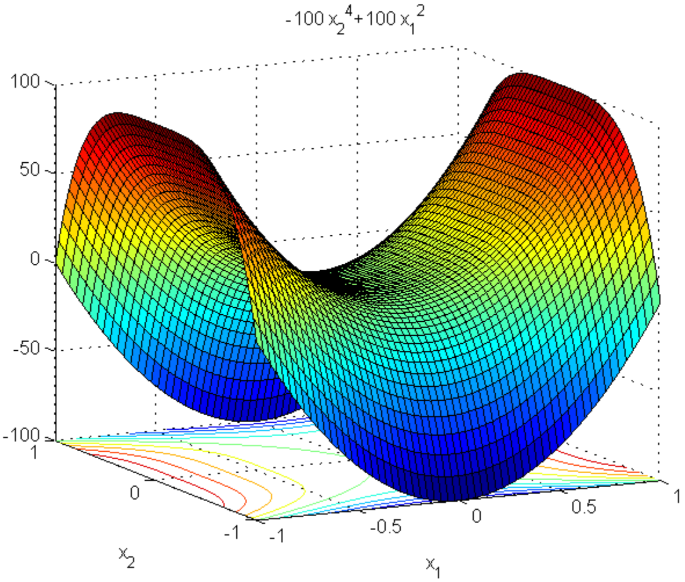
From second-order approximation at $\mathbf{x}_*$ we have:

$$f(\mathbf{x}) = f(\mathbf{x}_*) + \mathbf{g}_{\mathbf{x}_*}^T(\mathbf{x}-\mathbf{x}_*) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_*)^T\mathbf{H}_{\mathbf{x}_*}(\mathbf{x}-\mathbf{x}_*) + o(||\mathbf{x}-\mathbf{x}_*||^2). \quad (2)$$

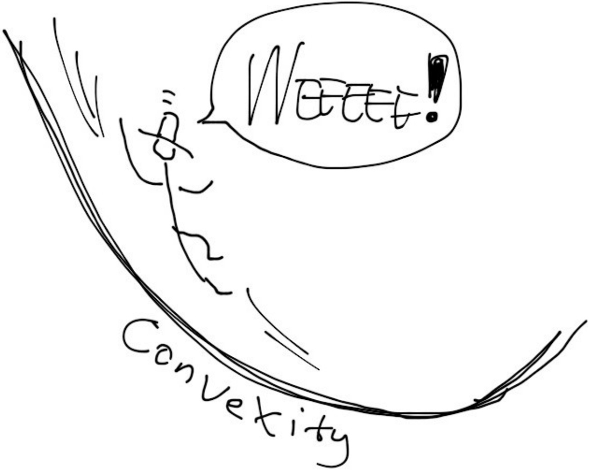
Let $\mathbf{x} = \mathbf{x}_* + t\mathbf{d}$, where $\mathbf{d}$ is a unit direction, i.e., $||\mathbf{d}|| = 1$. Using first-order necessary condition, and the fact that $\mathbf{x}_*$ is a local solution, we have

$$0 \leq \frac{f(\mathbf{x}_* + t\mathbf{d}) - f(\mathbf{x}_*)}{t^2} = \frac{1}{2}\mathbf{d}^T\mathbf{H}_{\mathbf{x}_*}\mathbf{d} + \frac{o(t^2)}{t^2}.$$

Taking $t \to 0$, we have $0 \leq \mathbf{d}^T\mathbf{H}_{\mathbf{x}_*}\mathbf{d}$. Since d is arbitrarily chosen, we have that $\mathbf{H}_{\mathbf{x}_*}$ is positive semi-definite.



The Hessian at the stationary point is positive semidefinite, but the stationary point is not a local minimum.


WEEEE!
Convexity

Convex sets and convex functions

Definition (convex set)

A set $\mathcal{S} \in \mathbb{R}^n$ is convex if and only if, for every point $\mathbf{x}_1$, $\mathbf{x}_2$ in $\mathcal{S}$, the point

$$\mathbf{x}(\lambda) = \lambda \mathbf{x}_2 + (1 - \lambda)\mathbf{x}_1, \quad 0 \le \lambda \le 1$$

also belongs to the set.

Convex sets and convex functions

Definition (convex function)

A function $f : \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \in \mathbb{R}^n$ defined on a nonempty convex set $\mathcal{X}$ is called convex on $\mathcal{X}$ if and only if, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$:

$$f(\lambda \mathbf{x}_2 + (1-\lambda)\mathbf{x}_1) \leq \lambda f(\mathbf{x}_2) + (1-\lambda) f(\mathbf{x}_1),$$

where $0 \leq \lambda \leq 1$.

Exercise 4: Show the intersection of convex sets is convex; Show $f_1 + f_2$ is convex on the set $\mathcal{S}$ if f_1, f_2 are convex on $\mathcal{S}$.

Exercise 5: Show that $f(\mathbf{x}_1) \geq f(\mathbf{x}_0) + \mathbf{g}_{\mathbf{x}_0}^T (\mathbf{x}_1 - \mathbf{x}_0)$ for a convex function.

Convex sets and convex functions

A differentiable function is convex if and only if its Hessian is positive-semidefinite in its entire convex domain. (hint: use Taylor's theorem to have $f(\mathbf{x}_1) = f(\mathbf{x}_0) + \mathbf{g}_{\mathbf{x}_0}^T(\mathbf{x}_1 - \mathbf{x}_0) + 1/2(\mathbf{x}_1 - \mathbf{x}_0)^T H_{\mathbf{x}(\lambda)}(\mathbf{x}_1 - \mathbf{x}_0)$, for $\mathbf{x}(\lambda) = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_0$.)

A positive-definite Hessian implies *strict* convexity, but the converse is generally not true. Example?

first-order sufficient condition

If a differentiable convex function with a convex open domain has a stationary point, this point will be the global minimum. If the function is strictly convex, then the minimum will be unique.

If the function is convex but not strictly convex, will the minimum be unique?

If the function is strictly convex, will the minimum be not unique?

Gradient descent

In reality it is hard to solve for the optimal solution $\mathbf{x}_*$ by hand because (i) the system of equations from the first-order necessary condition may not be easy to solve or (ii) the objective may not have an analytical form. Therefore, we need an *iterative* procedure to produce a series $\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_k$ that converges to $\mathbf{x}_*$.

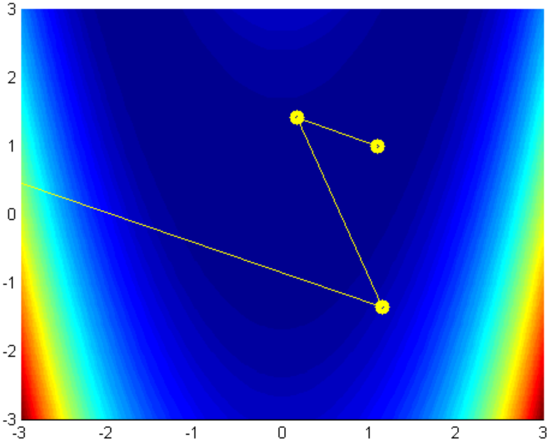
One naive way is to use the following: $\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{g}_k$. Why?

Exercise 6: Try this method for the following problem

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) = x_1^4 - 2x_1^2 x_2 + x_2^2$$

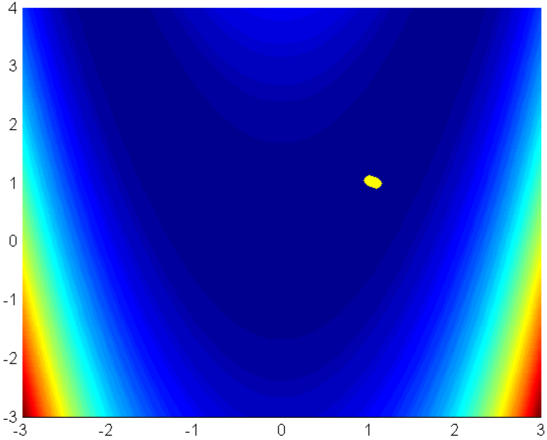
with $\mathbf{x}_0 = (1.1, 1)^T$. Explain your observation.

Gradient descent



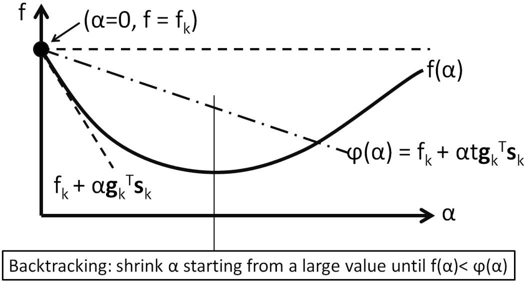
Results from Exercise 6. The gradient steps have correct directions but their step sizes are not desirable.

Gradient descent



Setting the step to 0.001 will allow the algorithm to converge but only slowly (takes more than 1000 steps to meet the target tolerance $||\mathbf{g}|| \leq 10^{-6}$)

Armijo line search



In Armijo line search, we construct a function

$$\phi(\alpha) = f_k + \alpha t \mathbf{g}_k^T \mathbf{s}_k,$$

for some $t \in [0.01, 0.3]$ and denote $f(\alpha) := f(\mathbf{x}_k + \alpha \mathbf{s}_k)$. Starting with a large value (default $\alpha = 1$), α is reduced by $\alpha = b\alpha$ for some $b \in [0.1, 0.8]$ until $f(\alpha) < \phi(\alpha)$, at which point it is guaranteed that $f(\alpha) < f_k$, since $\phi(\alpha) < f_k$ by nature.

Gradient algorithm with line search

Algorithm 1 Gradient algorithm

Select $\mathbf{x}_0$, $\varepsilon > 0$. Compute $\mathbf{g}_0$. Set $k = 0$.
while $||\mathbf{g}_k|| \geq \varepsilon$ **do**
    Compute $\alpha_k = \arg\min_{\alpha>0} f(\mathbf{x}_k - \alpha \mathbf{g}_k)$.
    Set $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k$.
end while

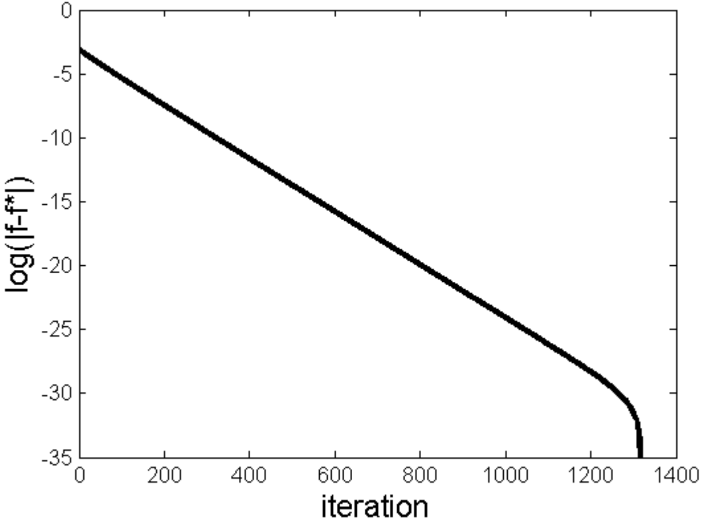


Figure: $\alpha = 0.001$, w/o line search

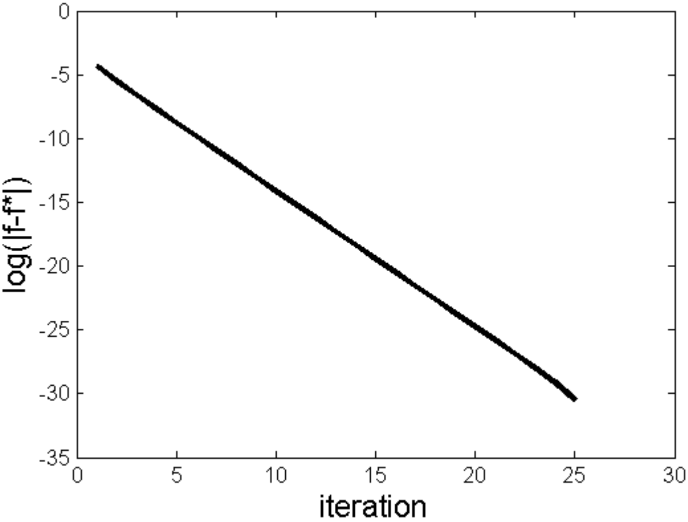


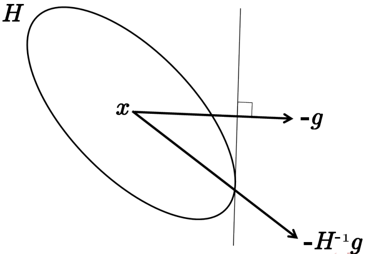
Figure: line search w/ $t = 0.3, b = 0.2$

Newton's method

Instead of using second-order approximation in line search, we can use it to find the direction.

$$f_{k+1} = f_k + \mathbf{g}_k \partial \mathbf{x}_k + \frac{1}{2} \partial \mathbf{x}_k^T \mathbf{H}_k \partial \mathbf{x}_k.$$

The first-order necessary condition for minimizing the approximated f_{k+1} requires $\mathbf{x}_{k+1} = \mathbf{x}_k - \mathbf{H}_k^{-1} \mathbf{g}_k$. If the function is locally strictly convex, this iteration will yield a lower function value. Newton's method will move efficiently in the neighborhood of a local minimum where local convexity is present.



Newton's method

Newton's method also requires line search since the second order approximation may not capture the actual function.

Algorithm 2 Newton's method

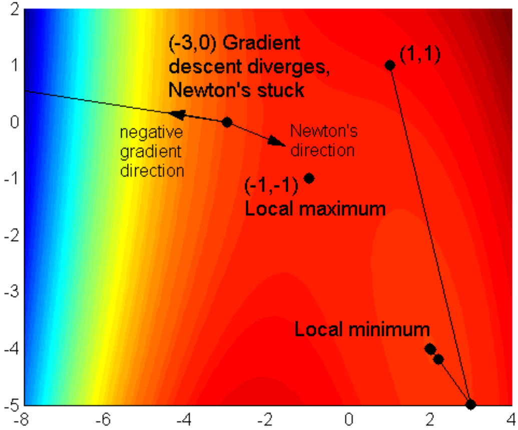
1: Select $\mathbf{x}_0$, $\varepsilon > 0$. Compute $\mathbf{g}_0$ and $\mathbf{H}_0$. Set $k = 0$.
2: **while** $||\mathbf{g}_k|| \geq \varepsilon$ **do**
3: Compute $\alpha_k = \arg\min_{\alpha>0} f(\mathbf{x}_k - \alpha \mathbf{H}_k^{-1}\mathbf{g}_k)$.
4: Set $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{H}_k^{-1}\mathbf{g}_k$.
5: **end while**

Exercise 7: Try this method for the following problem

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) = \frac{1}{3}x_1^3 + x_1 x_2 + \frac{1}{2}x_2^2 + 2x_2$$

with $\mathbf{x}_0 = (1,1)^T,\ (-1,-1)^T,\ (-3,0)^T$.

Newton's method



Exercise 7 cont.: Different starting points lead to different solutions. Newton's method does not guarantee a descent direction when the objective function is nonconvex.

Exercise (1/3)

4.6 Using the methods of this chapter, find the minimum of the function

$$f = (1 - x_1)^2 + 100(x_2 - x_1^2)^2.$$

This is the well-known Rosenbrock's "banana" function, a test function for numerical optimization algorithms.

4.8 Prove by completing the square that if a function $f(x_1, x_2)$ has a stationary point, then this point is

(a) a local minimum, if

$$\left(\partial^2 f/\partial x_1^2\right)\left(\partial^2 f/\partial x_2^2\right) - \left(\partial^2 f/\partial x_1 \partial x_2\right)^2 > 0 \quad \text{and} \quad \partial^2 f/\partial x_1^2 > 0;$$

(b) a local maximum, if

$$\left(\partial^2 f/\partial x_1^2\right)\left(\partial^2 f/\partial x_2^2\right) - \left(\partial^2 f/\partial x_1 \partial x_2\right)^2 > 0 \quad \text{and} \quad \partial^2 f/\partial x_1^2 < 0;$$

(c) a saddlepoint, if

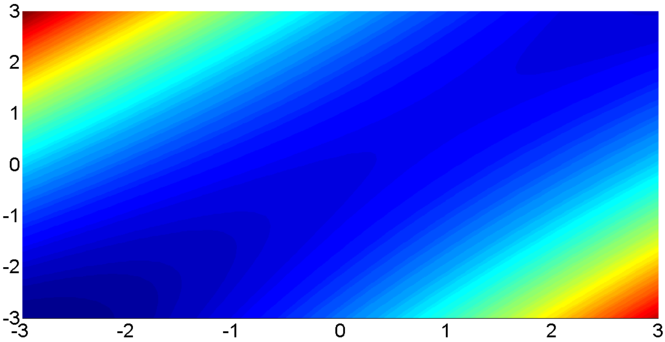
$$\left(\partial^2 f/\partial x_1^2\right)\left(\partial^2 f/\partial x_2^2\right) - \left(\partial^2 f/\partial x_1 \partial x_2\right)^2 < 0.$$

Exercise (2/3)

4.10 Show that the stationary point of the function

$$f = 2x_1^2 - 4x_1x_2 + 1.5x_2^2 + x_2$$

is a saddle. Find the directions of downslopes away from the saddle using the differential quadratic form.



Exercise (3/3)

4.12 Find the point in the plane $x_1 + 2x_2 + 3x_3 = 1$ in $\mathbb{R}^3$ that is nearest to the point $(-1, 0, 1)^T$.

4.15 Consider the function

$$f = -x_2 + 2x_1x_2 + x_1^2 + x_2^2 - 3x_1^2x_2 - 2x_1^3 + 2x_1^4.$$

(a) Show that the point $(1, 1)^T$ is stationary and that the Hessian is positive-semidefinite there.

(b) Find a straight line along which the second-order perturbation ∂f is zero.

Stabilization

Given a symmetric matrix $\mathbf{M}_k$, the gradient and Newton's methods can be classed together by the general iteration

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha \mathbf{M}_k \mathbf{g}_k,$$

where $\mathbf{M}_k = \mathbf{I}$ for gradient descent and $\mathbf{M}_k = \mathbf{H}_k^{-1}$ for Newton's method.

By the first-order Taylor approximation $f_{k+1} - f_k = -\alpha \mathbf{g}_k^T \mathbf{M}_k \mathbf{g}_k$, descent is accomplished for positive-definite $\mathbf{M}_k$.

To ensure descent, we can use $\mathbf{M}_k = (\mathbf{H}_k + \mu_k \mathbf{I})^{-1}$ and select a positive scalar μ_k.

Trust region (1/2)

Newton's method is faster if $\mathbf{x}_k$ is close to $\mathbf{x}_*$. If the search length $||\alpha \mathbf{H}_k^{-1}\mathbf{g}_k||$ is really short or if $\mathbf{H}_k$ is very different from the Hessian at $\mathbf{x}_*$, then $-\mathbf{g}_k$ may be a better direction to search.

Trust region algorithm: search using a quadratic approximation, but restrict the search step within the trust region with radius Δ at $\mathbf{x}_k$:

$$\begin{aligned} \min_{\mathbf{s}} \quad & f \approx \mathbf{g}_k^T \mathbf{s} + \frac{1}{2}\mathbf{s}^T \mathbf{H}_k \mathbf{s} \\ \text{subject to} \quad & ||\mathbf{s}|| \le \Delta \end{aligned}$$

If $||\mathbf{s}|| \le \Delta$ then

$$\mathbf{s}_k = -\mathbf{H}_k^{-1}\mathbf{g}_k \quad ||\mathbf{H}_k^{-1}\mathbf{g}_k|| < \Delta,$$

or if $||\mathbf{s}|| = \Delta$ then

$$\mathbf{s}_k = -(\mathbf{H}_k + \mu \mathbf{I})^{-1}\mathbf{g}_k, \mu > 0, \quad ||\mathbf{s}_k|| = \Delta.$$

Trust region (2/2)

Given Δ, we evaluate at $f(\mathbf{x}_k + \mathbf{s}_k)$ and calculate the ratio

$$r_k = \frac{f(\mathbf{x}_k) - f(\mathbf{x}_k + \mathbf{s}_k)}{\mathbf{g}_k^T \mathbf{s}_k + \frac{1}{2}\mathbf{s}_k^T \mathbf{H}_k \mathbf{s}_k}$$

The trust region radius is increased when $r_k > 0.75$ (indicating a "good step") and decreased when $r_k < 0.25$ (indicating a "bad step").

Algorithm 3 Trust region algorithm

1: Start with $\mathbf{x}_0$ and $\Delta_0 > 0$. Set $k = 0$.
2: Calculate the step $\mathbf{s}_k$.
3: Calculate the value $f(\mathbf{x}_k + \mathbf{s}_k)$ and the ratio r_k.
4: If $f(\mathbf{x}_k + \mathbf{s}_k) \geq f(\mathbf{x}_k)$, then set $\Delta_{k+1} = \Delta_k/2$, $\mathbf{x}_{k+1} = \mathbf{x}_k$, $k = k+1$ and go to Step 3.
5: Else, set $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$. If $r_k < 0.25$, then set $\Delta_{k+1} = \Delta_k/2$; if $r_k > 0.75$, then set $\Delta_{k+1} = 2\Delta_k$; otherwise set $\Delta_{k+1} = \Delta_k$. Set $k = k+1$ and go to Step 2.