Unconstrained Minimization (II)

Lecture 12, Convex Optimization

National Taiwan University

May 27, 2021

Table of contents

1. Steepest descent method
   - Steepest descent method
   - Examples for steepest descent methods
2. Newton's method in unconstrained problems
   - The Newton step
   - The Newton decrement
   - Newton's method
   - Examples

Steepest descent method (1/3)

- The first-order Taylor approximation of $f(x+v)$ around x is

$$f(x+v) \approx \hat{f}(x+v) = f(x) + \nabla f(x)^T v,$$

  where the term $\nabla f(x)^T v$ is the **directional derivative** of f at x in the direction v.
- It gives the approximate change in f for a small step v.
- The step v is a descent direction if the directional derivative is negative.

Steepest descent method (2/3)

- Let $||\cdot||$ be any norm on $\mathbf{R}^n$. We define a **normalized steepest descent direction** (with respect to the norm $||\cdot||$) as

$$\Delta x_{\mathrm{nsd}} = \arg\min_v \left\{ \nabla f(x)^T v \;\middle|\; ||v|| = 1 \right\},$$

which is a step of unit norm that gives the largest decrease in the linear approximation of f.

- It is convenient to consider a **steepest descent step** Δx_{sd} that is unnormalized, by scaling the normalized steepest descent direction in a particular way:

$$\Delta x_{\mathrm{sd}} = ||\nabla f(x)||_* \Delta x_{\mathrm{nsd}}.$$

Steepest descent method (3/3)

- Recall that for any given norm $||\cdot||$, the associated **dual norm**, denoted by $||\cdot||_*$, is defined as

$$||z||_* = \sup\left\{z^T x \mid ||x|| \leq 1\right\}.$$

- Note that for the steepest descent step, defined as

$$\Delta x_{\mathrm{sd}} = ||\nabla f(x)||_* \Delta x_{\mathrm{nsd}},$$

we have

$$\nabla f(x)^T \Delta x_{\mathrm{sd}} = ||\nabla f(x)||_* \nabla f(x)^T \Delta x_{\mathrm{nsd}} = -||\nabla f(x)||_*^2.$$

Steepest descent for Euclidean norm

- If we take the norm $||\cdot||$ to be the Euclidean norm, then the steepest descent direction is simply the negative gradient, i.e., $\Delta x_{\mathrm{sd}} = -\nabla f(x)$.
  - To see this, note that

$$\Delta x_{\mathrm{sd}} = ||\nabla f(x)||_2 \Delta x_{\mathrm{nsd}} = ||\nabla f(x)||_2 \left(-\frac{\nabla f(x)}{||\nabla f(x)||_2} \right).$$

- The steepest descent method for the Euclidean norm coincides with the gradient descent method.

Steepest descent Algorithm

- The steepest descent method uses the steepest descent direction as search direction.

- **Algorithm 4.** Steepest descent method.
  given a starting point $x \in \mathbf{dom}\ f$.
  repeat
  1. Compute steepest descent direction Δx_{sd}.
  2. *Line search*. Choose t via backtracking or exact line search.
  3. *Update*. $x := x + t\Delta x_{\mathrm{sd}}$.

  until stopping criterion is satisfied.

- When exact line search is used, scale factors in the descent direction have no effect, so the normalized or unnormalized direction can be used.

Steepest descent for quadratic norm (1/2)

- We consider the quadratic norm

$$||z||_P = (z^T P z)^{1/2} = ||P^{1/2} z||_2,$$

where $P \in \mathbf{S}^n_{++}$.

- The normalized steepest descent direction is given by

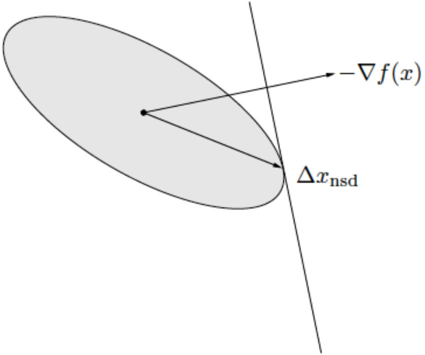
$$\Delta x_{\mathrm{nsd}} = -\left(\nabla f(x)^T P^{-1} \nabla f(x)\right)^{-1/2} P^{-1} \nabla f(x).$$

- The dual norm is given by $||z||_* = ||P^{-1/2} z||_2$, so the steepest descent step with respect to $||\cdot||_P$ is given by

$$\Delta x_{\mathrm{sd}} = -P^{-1} \nabla f(x).$$

Steepest descent for quadratic norm (2/2)

- The normalized steepest descent direction for a quadratic norm is illustrated in the following figure.



Choice of norm for steepest descent

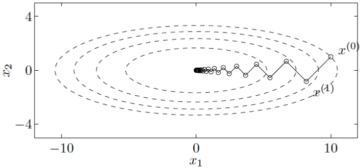
- The choice of norm used to define the steepest descent direction can have a dramatic effect on the convergence rate.
- We consider the case of steepest descent with quadratic P-norm.
- Recall that the steepest descent method with quadratic P-norm is the same as the gradient method applied to the problem after the change of coordinates $\bar{x} = P^{1/2}x$.
- We know that the gradient method works well when the condition numbers of the sublevel sets (or the Hessian near the optimal point) are moderate, and works poorly when the condition numbers are large.

Revisit of the Example of a quadratic problem in $\mathbf{R}^2$

- We consider again the simple example with the quadratic objective function on $\mathbf{R}^2$

$$f(x) = \frac{1}{2}(x_1^2 + \gamma x_2^2), \qquad \gamma > 0.$$

- The following figure illustrates the gradient descent method with exact line search for the case $\gamma = 10$.



- What if we use the steepest descent method with the choice of the quadratic P-norm where $P = \begin{bmatrix} 1 & 0 \\ 0 & \gamma \end{bmatrix}$?

Examples (1/7)

- We illustrate some of these ideas using the nonquadratic problem in $\mathbf{R}^2$ with objective function

$$f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}.$$

- We apply the steepest descent method to the problem, using the two quadratic norms defined by

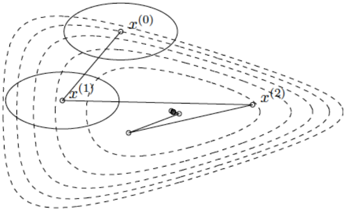
$$P_1 = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}, P_2 = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}.$$

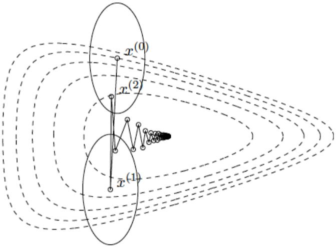
- In both cases we use a backtracking line search with $\alpha = 0.1$ and $\beta = 0.7$.

Examples (2/7)

- The following figures show the iterates for steepest descent with norm $||\cdot||_{P_1}$ and norm $||\cdot||_{P_2}$, respectively.

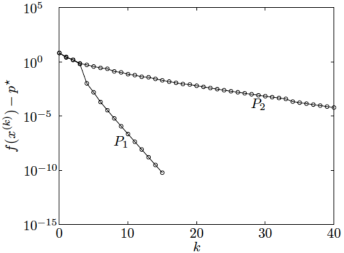
Examples (3/7)





Examples (4/7)

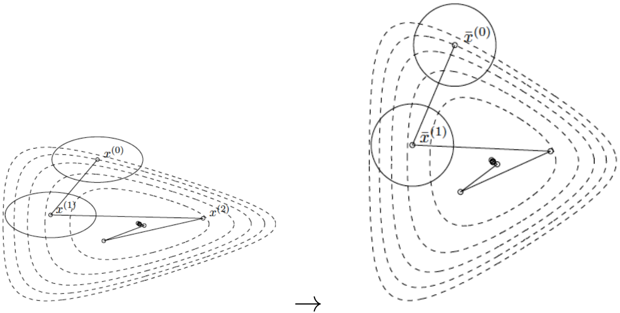
- The following figure shows the error versus iteration number for both norms and shows that the choice of norm strongly influences the convergence.
- With the norm $||\cdot||_{P_1}$, the convergence is a bit more rapid than the gradient method, whereas with the norm $||\cdot||_{P_2}$, the convergence is far slower.



Examples (5/7)

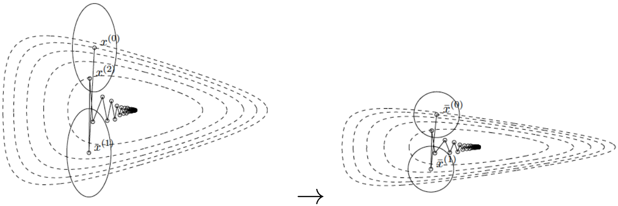
- This can be explained by examining the problems after the changes of coordinates $\bar{x} = P_1^{1/2} x$ and $\bar{x} = P_2^{1/2} x$, respectively.
- The change of variables associated with P_1 yields sublevel sets with a modest condition number, so the convergence is fast.

Examples (6/7)



- The change of variables associated with P_2 yields sublevel sets that are more poorly conditioned, which explains the slower convergence.

Examples (7/7)



Choosing the "Best" Quadratic P-Norm

- From the previous discussions, we found that the choice of $P \in \mathbf{S}^n_{++}$ for the quadratic P-norm applied in the steepest descent method affects the convergence rate quite a lot.
- Question: can we always choose the "best" matrix P for the quadratic P-norm in the steepest descent method?

The Newton step

Newton step

For $x \in \textbf{dom}\ f$, the vector

$$\Delta x_{\mathrm{nt}} = -\nabla^2 f(x)^{-1} \nabla f(x)$$

is called the **Newton step** (for f, at x).

- If $\nabla^2 f(x)$ is positive definite, it implies that

$$\nabla f(x)^T \Delta x_{\mathrm{nt}} = -\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x) < 0$$

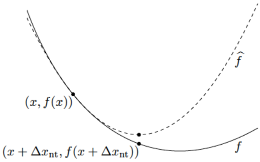
  unless $\nabla f(x) = 0$, so the Newton step is a descent direction (unless x is optimal).
- The Newton step can be interpreted and motivated in several ways.

Minimizer of second-order approximation (1/2)

- The second-order Taylor approximation $\hat{f}$ of f at x is

$$\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v,$$

  which is a convex quadratic function of v, and is minimized when $v = \Delta x_{\mathrm{nt}}$.
- Thus, the Newton step Δx_{nt} is what should be added to the point x to minimize the second-order approximation of f at x.



Minimizer of second-order approximation (2/2)

- If the function f is quadratic, then $x + \Delta x_{\mathrm{nt}}$ is the exact minimizer of f.

Steepest descent direction in Hessian norm (1/2)

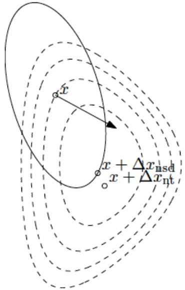
- The Newton step is also the steepest descent direction at x, for the quadratic norm defined by the Hessian $\nabla^2 f(x)$, i.e.,

$$||u||_{\nabla^2 f(x)} = (u^T \nabla^2 f(x) u)^{1/2}.$$

- This gives another insight into why the Newton step should be a good search direction, and a very good search direction when x is near x^*.
- Recall that steepest descent, with quadratic norm $|| \cdot ||_P$, converges very rapidly when the Hessian, after the associated change of coordinates, has small condition number.
- In particular, near x^*, a very good choice is $P = \nabla^2 f(x^*)$.

Steepest descent direction in Hessian norm (2/2)

- When x is near x^*, we have $\nabla^2 f(x) \approx \nabla^2 f(x^*)$, which explains why the Newton step is a very good choice of search direction.



- In the above figure, the arrow denotes the gradient descent direction.

Solution of linearized optimality condition (1/3)

- We can linearize the optimality condition $\nabla f(x^*) = 0$ near x and obtain

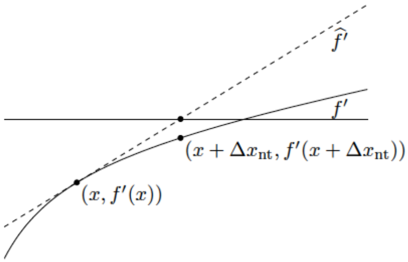
$$\nabla f(x+v) \approx \nabla f(x) + \nabla^2 f(x) v = 0,$$

  which is a linear equation in v, with solution $v = \Delta x_{\mathrm{nt}}$.
- So the Newton step Δx_{nt} is what must be added to x so that the linearized optimality condition holds.
- This suggests that when x is near x^* (so the optimality conditions almost hold), the update $x + \Delta x_{\mathrm{nt}}$ should be a very good approximation of x^*.
- When $n = 1$, i.e., $f : \mathbf{R} \to \mathbf{R}$, this interpretation is particularly simple.

Solution of linearized optimality condition (2/3)

- The solution x^* of the minimization problem is characterized by $f'(x^*) = 0$, i.e., it is the zero-crossing of the derivative f', which is monotonically increasing since f is strongly convex.
- Given our current approximation x of the solution, we form a first-order Taylor approximation of f' at x.
- The zero-crossing of this affine approximation is then $x + \Delta x_{\mathrm{nt}}$.

Solution of linearized optimality condition (3/3)



Affine invariance of the Newton step

- An important feature of the Newton step is that it is independent of linear (or affine) changes of coordinates.

Affine invariance of the Newton step

- An important feature of the Newton step is that it is independent of linear (or affine) changes of coordinates.
- Suppose $T \in \mathbf{R}^{n \times n}$ is nonsingular, and define $\bar{f}(y) = f(Ty)$. Then we have $\nabla \bar{f}(y) = T^T \nabla f(x), \nabla^2 \bar{f}(y) = T^T \nabla^2 f(x) T$, where $x = Ty$.

Affine invariance of the Newton step

- An important feature of the Newton step is that it is independent of linear (or affine) changes of coordinates.
- Suppose $T \in \mathbf{R}^{n \times n}$ is nonsingular, and define $\bar{f}(y) = f(Ty)$. Then we have $\nabla \bar{f}(y) = T^T \nabla f(x), \nabla^2 \bar{f}(y) = T^T \nabla^2 f(x) T$, where $x = Ty$.
- The Newton step for $\bar{f}$ at y is therefore

$$\begin{aligned} \Delta y_{nt} &= -(T^T \nabla^2 f(x) T)^{-1} (T^T \nabla f(x)) \\ &= -T^{-1} \nabla^2 f(x)^{-1} \nabla f(x) \\ &= T^{-1} \Delta x_{\mathrm{nt}}, \end{aligned}$$

where Δx_{nt} is the Newton step for f at x. Hence the Newton steps of f and $\bar{f}$ are related by the same linear transformation.

$$x + \Delta x_{\mathrm{nt}} = T(y + \Delta y_{nt}).$$

The Newton decrement (1/2)

- The quantity

$$\lambda(x) = \left(\nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x)\right)^{1/2}$$

is called the **Newton decrement** at x.

- We can relate the Newton decrement to the quantity

$$f(x) - \inf_y \hat{f}(y),$$

where $\hat{f}$ is the second-order approximation of f at x:[1]

$$f(x) - \inf_y \hat{f}(y) = f(x) - \hat{f}(x + \Delta x_{\mathrm{nt}}) = \frac{1}{2}\lambda(x)^2.$$

[1] Note: $\hat{f}(x+v) = f(x) + \nabla f(x)^T v + \frac{1}{2} v^T \nabla^2 f(x) v$.

The Newton decrement (2/2)

- Thus, $\lambda^2/2$ is an estimate of $f(x) - p^*$, based on the quadratic approximation of f at x.
- We can also express the Newton decrement as

$$\lambda(x) = \left(\Delta x_{nt}^T \nabla^2 f(x) \Delta x_{\mathrm{nt}}\right)^{1/2},$$

which shows that λ is the norm of the Newton step, in the quadratic norm defined by the Hessian, i.e., the norm

$$||u||_{\nabla^2 f(x)} = \left(u^T \nabla^2 f(x) u\right)^{1/2}.$$

- The Newton decrement is, like the Newton step, affine invariant: the Newton decrement of $\bar{f}(y) = f(Ty)$ at y, where T is nonsingular, is the same as the Newton decrement of f at $x = Ty$.

Newton's method

- Newton's method, as outlined below, is sometimes called the **damped Newton method**, to distinguish it from the pure Newton method, which uses a fixed step size $t = 1$.
- **Algorithm 5.** (Damped) Newton's method.
  given a starting point $x \in \textbf{dom } f$, tolerance $\epsilon > 0$.
  repeat
  1. *Compute the Newton step and decrement.*
  $$\Delta x_{\rm nt} := -\nabla^2 f(x)^{-1} \nabla f(x); \lambda^2 := \nabla f(x)^T \nabla^2 f(x)^{-1} \nabla f(x).$$
  2. *Stopping criterion.* **quit** if $\lambda^2/2 \leq \epsilon$.
  3. *Line search.* Choose step size t by backtracking line search.
  4. *Update.* $x := x + t\Delta x_{\rm nt}$.
- This is essentially the general descent method using the Newton step as search direction.

Convergence analysis (1/3)

- We assume, as before, that f is twice continuously differentiable, and strongly convex with constant m, i.e., $\nabla^2 f(x) \succeq mI$ for $x \in S$. This implies that there exists an $M > 0$ such that $\nabla^2 f(x) \preceq MI$ for all $x \in S$.
- In addition, we assume that the Hessian of f is **Lipschitz continuous** on S with constant L, i.e.,

$$||\nabla^2 f(x) - \nabla^2 f(y)||_2 \leq L||x - y||_2$$

  for all $x, y \in S$.
- The coefficient L, which can be interpreted as a bound on the third derivative of f, can be taken to be zero for a quadratic function.

Convergence analysis (2/3)

- More generally L measures how well f can be approximated by a quadratic model, so we can expect the **Lipschitz constant** L to play a critical role in the performance of Newton's method.
- Intuition suggests that Newton's method will work very well for a function whose quadratic model varies slowly (i.e., has small L).
- It can be shown that there are numbers η and γ with $0 < \eta \leq m^2/L$ and $\gamma > 0$ such that the following hold.
    - If $||\nabla f(x^{(k)})||_2 \geq \eta$, then

$$f(x^{(k+1)}) - f(x^{(k)}) \leq -\gamma.$$

Convergence analysis (3/3)

- If $||\nabla f(x^{(k)})||_2 < \eta$, then the backtracking line search selects $t^{(k)} = 1$ and

$$\frac{L}{2m^2}||\nabla f(x^{(k+1)})||_2 \leq \left(\frac{L}{2m^2}||\nabla f(x^{(k)})||_2\right)^2.$$

- The case when $||\nabla f(x^{(k)})||_2 \geq \eta$ is referred to as the **damped Newton phase**; the case when $||\nabla f(x^{(k)})||_2 < \eta$ is called the **quadratically convergent stage**.
- The number of iterations needed is bounded above by

$$6 + \frac{M^2 L^2 / m^5}{\alpha\beta \min\{1, 9(1-2\alpha)^2\}}(f(x^{(0)}) - p^*).$$

- The proof is omitted here. Interested audience can refer to the textbook.

Example in $\mathbf{R}$ (1/2)

- Consider the unconstrained problem with an objective function $f_0 : \mathbf{R} \to \mathbf{R}, \mathbf{dom}\ f_0 = \mathbf{R}_{++}$,

$$f_0(x) = x^3 - 6x.$$

- Then, $f_0'(x) = 3x^2 - 6$ and $f_0''(x) = 6x$.
- The Newton step is

$$\Delta x_{nt} = -\frac{f_0'(x)}{f_0''(x)} = -\frac{x}{2} + \frac{1}{x}.$$

- The optimal point can be shown to be $x^* = \sqrt{2} = 1.414213562373095\ldots$

Example in $\mathbf{R}$ (2/2)

- It is observed that f_0 is convex in **dom** f_0.
- Within the interval $(1, 3)$, f_0 is strictly convex and satisifes $m \leq f_0''(x) \leq M$ where $m = 6$ and $M = 18$.
- The Lipschitz constant can be chosen as $L = 6$.

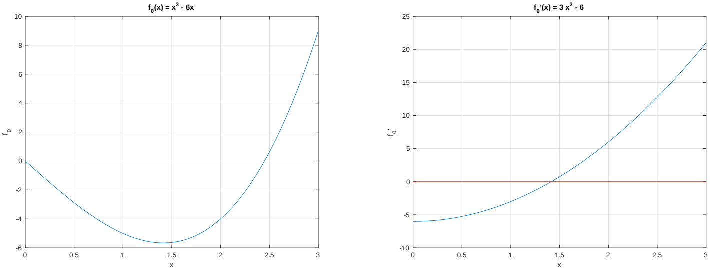


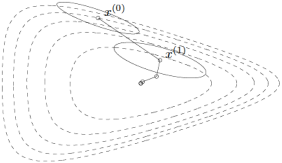
Figure 1: Left: $f_0(x) = x^3 - 6x$. Right $f_0'(x) = 3x^2 - 6$.

Example in $\mathbf{R}^2$ (1/3)

- We apply Newton's method with backtracking line search, with parameters $\alpha = 0.1, \beta = 0.7$, on the test function $f(x_1, x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$.
- The next figure shows the Newton iterates and the ellipsoids

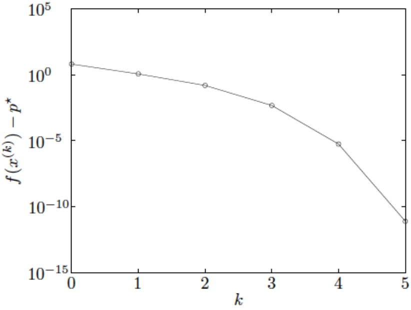
$$\left\{ x \mid \|x - x^{(k)}\|_{\nabla^2 f(x^{(k)})} \leq 1 \right\}$$

for the first two iterates $k = 0, 1$.



Example in $\mathbf{R}^2$ (2/3)

- The method works well because these ellipsoids give good approximations of the shape of the sublevel sets.
- The error versus iteration number for the same example is shown below:



Example in $\mathbf{R}^2$ (3/3)

- This plot shows that convergence to a very high accuracy is achieved in only five iterations.
- Quadratic convergence is clearly apparent: The last step reduces the error from about 10^{-5} to 10^{-10}.

Summary (1/2)

- Newton's method has several very strong advantages over gradient and steepest descent methods:
  - Convergence of Newton's method is rapid in general, and quadratic near x^*. Once the quadratic convergence phase is reached, at most six or so iterations are required to produce a solution of very high accuracy.
  - Newton's method is affine invariant.
  - It is insensitive to the choice of coordinates, or the condition number of the sublevel sets of the objective.
  - Newton's method scales well with problem size.
  - Its performance on problems in $\mathbf{R}^{10000}$ is similar to its performance on problems in $\mathbf{R}^{10}$, with only a modest increase in the number of steps required.
  - The good performance of Newton's method is not dependent on the choice of algorithm parameters.

Summary (2/2)

- In contrast, the choice of norm for steepest descent plays a critical role in its performance.
- The main disadvantage of Newton's method is the cost of forming and storing the Hessian, and the cost of computing the Newton step, which requires solving a set of linear equations.