Unconstrained Minimization (I)

Lecture 11, Convex Optimization

National Taiwan University

May 20, 2021

Table of contents

Unconstrained Minimization Problems

- In this chapter, we discuss methods for solving the unconstrained optimization problem

$$\text{minimize } f(x)$$

  where $f : \mathbf{R}^n \to \mathbf{R}$ is convex and twice continuously differentiable (which implies that $\mathbf{dom}\ f$ is open).
- We will assume that the problem is solvable, i.e., there exists an optimal point x^*.
- We denote the optimal value, $\inf_x f(x) = f(x^*)$, as p^*.
- Since f is differentiable and convex, a necessary and sufficient condition for a point x^* to be optimal is $\nabla f(x^*) = 0$.

Solving Unconstrained Minimization Problems

- Thus, solving the unconstrained minimization problem

$$\text{minimize } f(x)$$

  is the same as finding a solution of

$$\nabla f(x^*) = 0,$$

  which is a set of n equations in the n variables $x_1, \ldots, x_n$.
- We sometimes can find an analytical solution for $\nabla f(x^*) = 0$, but in general it must be solved by an iterative algorithm that computes a sequence of points $x^{(0)}, x^{(1)}, \ldots \in \textbf{dom } f$ with $f(x^{(k)}) \to p^*$ as $k \to \infty$.
- Such a sequence of points is called a **minimizing sequence** for the problem "minimize $f(x)$."
- The algorithm is terminated when $f(x^{(k)}) - p^* \leq \epsilon$, where $\epsilon > 0$ is some specified tolerance.

Example 1 – Quadratic minimization and least-squares

- The general convex quadratic minimization problem has the form
$$\text{minimize } \frac{1}{2}x^T P x + q^T x + r,$$
where $P \in \mathbf{S}^n_+$, $q \in \mathbf{R}^n$, and $r \in \mathbf{R}$.
- This problem can be solved via the optimality conditions,
$$Px^* + q = 0.$$
- When $P \succ 0$, there is a unique solution, $x^* = -P^{-1}q$.
- In the case when $P \notin \mathbf{S}^n_{++}$,
  1. any solution of $Px^* = -q$ is optimal (if a solution exists);
  2. if $Px^* = -q$ does not have a solution, then the problem is unbounded below.

Example 2 – Unconstrained geometric programming

- As a second example, we consider an unconstrained geometric program in convex form,

$$\text{minimize } f(x) = \log \sum_{i=1}^{m} \exp(a_i^T x + b_i).$$

- The optimality condition is

$$\nabla f(x^*) = \frac{1}{\sum_{j=1}^{m} \exp(a_j^T x^* + b_j)} \sum_{i=1}^{m} \exp(a_i^T x^* + b_i) a_i = 0,$$

which in general has no analytical solution, so here we must resort to an iterative algorithm.

- Since **dom** $f = \mathbf{R}^n$ for this problem, any point can be chosen as the initial point $x^{(0)}$.

Example 3 – Analytic center of linear inequalities (1/2)

- We consider the optimization problem

$$\text{minimize } f(x) = -\sum_{i=1}^{m} \log(b_i - a_i^T x),$$

where the domain of f is the open set

$$\mathbf{dom}\ f = \left\{ x \mid a_i^T x < b_i, \quad i = 1, \ldots, m \right\}.$$

- The objective function f in this problem is called the **logarithmic barrier** for the inequalities $a_i^T x \leq b_i$.
- The solution of the problem, if it exists, is called the **analytic center** of the inequalities.

Example 3 – Analytic center of linear inequalities (2/2)

- The initial point $x^{(0)}$ must satisfy the strict inequalities

$$a_i^T x^{(0)} < b_i, i = 1, ..., m.$$

- Since f is closed[1], the sublevel set S for any such point is closed.

[1] A function $f : \mathbf{R}^n \to \mathbf{R}$ is said to be **closed** if the sublevels $\{x \in \mathbf{dom}\ f \mid f(x) \leq \alpha\}$ is closed for each $\alpha \in \mathbf{R}$. It is equivalent to say that the epigraph of f, **epi** f, is closed.

Strong convexity (1/2)

- We assume that the objective function is **strongly convex** on S: there exists an $m > 0$ such that

$$\nabla^2 f(x) \succeq mI$$

for all $x \in S$.

- If f is strongly convex, then for $x, y \in S$, there exists some z on the line segment $[x, y]$ such that

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 f(z)(y - x) \\ &\geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2}||y - x||_2^2. \end{aligned}$$

- When $m = 0$, we recover the basic inequality characterizing convexity; for $m > 0$ we obtain a better lower bound on $f(y)$ than follows from convexity alone.

Strong convexity (2/2)

- Note that $f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}||y-x||_2^2$ can be minimized by $\tilde{y} = x - (1/m)\nabla f(x)$.
- Therefore we have

$$
\begin{aligned}
f(y) &\geq f(x) + \nabla f(x)^T(y-x) + \frac{m}{2}||y-x||_2^2 \\
&\geq f(x) + \nabla f(x)^T(\tilde{y}-x) + \frac{m}{2}||\tilde{y}-x||_2^2 \\
&= f(x) - \frac{1}{2m}||\nabla f(x)||_2^2.
\end{aligned}
$$

- Since this holds for any $y \in S$, we have

$$p^* \geq f(x) - \frac{1}{2m}||\nabla f(x)||_2^2.$$

Strong convexity and implications (1/2)

- This inequality shows that if the gradient $||\nabla f(x)||_2$ is small at some point x, then x is nearly optimal. Specifically,

$$||\nabla f(x)||_2 \leq (2m\epsilon)^{1/2} \Longrightarrow f(x) - p^* \leq \epsilon.$$

- We can also derive a bound on $||x - x^*||_2$, the distance between x and any optimal point x^*, in terms of $||\nabla f(x)||_2$:

$$||x - x^*||_2 \leq \frac{2}{m}||\nabla f(x)||_2.$$

- One consequence of the above inequality is that the optimal point x^* is unique.

Strong convexity and implications (2/2)

- Proof of the inequality $||x - x^*||_2 \leq \frac{2}{m}||\nabla f(x)||_2$:

$$\begin{aligned} p^* = f(x^*) &\geq f(x) + \nabla f(x)^T(x^* - x) + \frac{m}{2}||x^* - x||_2^2 \\ &\geq f(x) - ||\nabla f(x)||_2 \; ||x^* - x||_2 + \frac{m}{2}||x^* - x||_2^2, \end{aligned}$$

where we use the Cauchy-Schwarz inequality in the second inequality. Since $p^* \leq f(x)$, we must have

$$-||\nabla f(x)||_2 \; ||x^* - x||_2 + \frac{m}{2}||x^* - x||_2^2 \leq 0. \text{ (QED)}$$

Upper bound on $\nabla^2 f(x)$ (1/2)

- The inequality

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{m}{2} ||y - x||_2^2$$

  implies that the sublevel sets contained in S are bounded.
- Therefore, the maximum eigenvalue of $\nabla^2 f(x)$, which is a continuous function of x on S, is bounded above on S, i.e., there exists a constant M such that

$$\nabla^2 f(x) \preceq MI$$

  for all $x \in S$.

Upper bound on $\nabla^2 f(x)$ (2/2)

- This upper bound on the Hessian implies for any $x, y \in S$,

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{M}{2} ||y - x||_2^2.$$

- Minimizing each side over y yields

$$p^* \leq f(x) - \frac{1}{2M} ||\nabla f(x)||_2^2.$$

Condition number of convex sets (1/3)

- From the above discussions, we have

$$mI \preceq \nabla^2 f(x) \preceq MI$$

  for all $x \in S$.
- The ratio $\kappa = M/m$ is thus an upper bound on the **condition number** of the matrix $\nabla^2 f(x)$, i.e., the ratio of its largest eigenvalue to its smallest eigenvalue.
- We define the **width** of a convex set $C \subseteq \mathbf{R}^n$, in the direction q, where $||q||_2 = 1$, as

$$W(C, q) = \sup_{z \in C} q^T z - \inf_{z \in C} q^T z.$$

Condition number of convex sets (2/3)

- The **minimum width** and **maximum width** of C are given by

$$W_{min} = \inf_{||q||_2=1} W(C, q), W_{max} = \sup_{||q||_2=1} W(C, q).$$

- The **condition number** of the convex set C is defined as

$$\mathbf{cond}(C) = \frac{W_{max}^2}{W_{min}^2},$$

  i.e., the square of the ratio of its maximum width to its minimum width.

- The condition number of C gives a measure of its anisotropy or eccentricity[2].

[2] Refer to anisotropy and eccentricity for their definitions.

Condition number of convex sets (3/3)

- If the condition number of a set C is small (say, near one) it means that the set has approximately the same width in all directions, i.e., it is nearly spherical.
- If the condition number is large, it means that the set is far wider in some directions than in others.

Example – Condition number of an ellipsoid (1/2)

- Let $\mathcal{E}$ be the ellipsoid

$$\mathcal{E} = \left\{ x \;\middle|\; (x - x_0)^T A^{-1} (x - x_0) \leq 1 \right\},$$

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

- The width of $\mathcal{E}$ in the direction q, where $||q||_2 = 1$, is

$$\begin{aligned} \sup_{z \in \mathcal{E}} q^T z - \inf_{z \in \mathcal{E}} q^T z &= (||A^{1/2} q||_2 + q^T x_0) - (-||A^{1/2} q||_2 + q^T x_0) \\ &= 2||A^{1/2} q||_2. \end{aligned}$$

Example – Condition number of an ellipsoid (2/2)

- So, the minimum and maximum width of $\mathcal{E}$ are

$$W_{min} = 2\lambda_{min}(A)^{1/2}, W_{max} = 2\lambda_{max}(A)^{1/2},$$

  and the condition number is

$$\textbf{cond}(\mathcal{E}) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} = \kappa(A),$$

  where $\kappa(A)$ denotes the condition number of the matrix A, i.e., the ratio of its maximum singular value to its minimum singular value.
- Thus the condition number of the ellipsoid $\mathcal{E}$ is the same as the condition number of the matrix A that defines it.

Descent methods (1/3)

- The algorithms described in this chapter produce a minimizing sequence $x^{(k)}$, $k = 0, 1, ...,$ where

$$x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$$

and $t^{(k)} > 0$ (except when $x^{(k)}$ is optimal).

- The vector $\Delta x^{(k)} \in \mathbf{R}^n$ is called the step or search direction, and $k = 0, 1, \ldots$ denotes the iteration number.
- The scalar $t^{(k)} \geq 0$ is called the step size or step length at iteration k (even though it is not equal to $||x^{(k+1)} - x^{(k)}||$ unless $||\Delta x^{(k)}|| = 1$).

Descent methods (2/3)

- When we focus on one iteration of an algorithm, we sometimes drop the superscripts and use the lighter notation

$$x^+ = x + t\Delta x, \text{ or } x := x + t\Delta x,$$

in place of

$$x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}.$$

- All the methods we study are descent methods, which means that

$$f(x^{(k+1)}) < f(x^{(k)}),$$

except when $x^{(k)}$ is optimal.

- This implies that for all k we have $x^{(k)} \in S$, the initial sublevel set, and in particular we have $x^{(k)} \in \textbf{dom } f$.

Descent methods (3/3)

- From convexity we know that $\nabla f(x^{(k)})^T(y - x^{(k)}) \geq 0$ implies $f(y) \geq f(x^{(k)})$, so the search direction in a descent method must satisfy $\nabla f(x^{(k)})^T \Delta x^{(k)} < 0$, i.e., it must make an acute angle with the negative gradient.
- We call such a direction a **descent direction** (for f, at $x^{(k)}$).

General descent method (1/2)

- The outline of a general descent method is as follows, which alternates between two steps: determining a descent direction Δx, and the selection of a step size t.

- **Algorithm 1.** General descent method.
  given a starting point $x \in \textbf{dom } f$.
  repeat
  1. Determine a descent direction Δx.
  2. *Line search*. Choose a step size $t > 0$.
  3. *Update*. $x := x + t\Delta x$.

  until stopping criterion is satisfied.

General descent method (2/2)

- The second step is called the **line search** (or ray search, to be more accurate) since selection of the step size t determines where along the line $\{x + t\Delta x \mid t \in \mathbf{R}_+\}$ the next iterate will be.
- A practical descent method has the same general structure, but might be organized differently.
    - For example, the stopping criterion is often checked while, or immediately after, the descent direction Δx is computed.
    - The stopping criterion is often of the form $||\nabla f(x)||_2 \leq \eta$, where η is small and positive, as suggested by the suboptimality condition

$$p^* \geq f(x) - \frac{1}{2m}||\nabla f(x)||_2^2.$$

Exact line search

- One line search method sometimes used in practice is **exact line search**, in which t is chosen to minimize f along the ray $\{x + t\Delta x \mid t \geq 0\}$:

$$t = \arg\min_{s \geq 0} f(x + s\Delta x).$$

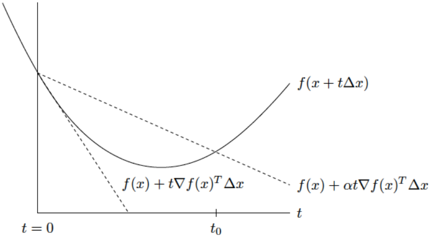
- An exact line search is used when the cost of the minimization problem with one variable is low compared to the cost of computing the search direction itself.

Backtracking line search (1/5)

- Most line searches used in practice are inexact: the step length is chosen to approximately minimize f along the ray $\{x + t\Delta x \mid t \geq 0\}$, or even to just reduce f 'enough'.
- Many inexact line search methods have been proposed. We study here one of them, called **backtracking line search**, which is very simple and quite effective.
- It depends on two constants α, β with $0 < \alpha < 0.5,\ 0 < \beta < 1$.

Backtracking line search (2/5)

- **Algorithm 2.** Backtracking line search.
  given a descent direction Δx for f at $x \in \textbf{dom } f$,
  $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$.
  $t := 1$.
  while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$,
  $t := \beta t$.



Backtracking line search (3/5)

- Since Δx is a descent direction, we have $\nabla f(x)^T \Delta x < 0$, so for small enough t we have
$$f(x + t\Delta x) \approx f(x) + t\nabla f(x)^T \Delta x < f(x) + \alpha t \nabla f(x)^T \Delta x,$$
which shows that the backtracking line search eventually terminates.
- The constant α can be interpreted as the fraction of the decrease in f predicted by linear extrapolation that we will accept.
- This figure suggests, and it can be shown, that the backtracking exit inequality $f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$ holds for $t \geq 0$ in an interval $(0, t_0]$, where t_0 is the only positive value that satisfies
$$f(x + t_0 \Delta x) = f(x) + \alpha t_0 \nabla f(x)^T \Delta x.$$

Backtracking line search (4/5)

- It follows that the backtracking line search stops with a step length t that satisfies $t = 1$, or $t \in (\beta t_0, t_0]$.
- The first case occurs when the step length $t = 1$ satisfies the backtracking condition, i.e., $1 \leq t_0$.
- In particular, we can say that the step length obtained by backtracking line search satisfies

$$t \geq \min\{1, \beta t_0\}.$$

- When **dom** f is not all of $\mathbf{R}^n$, the condition $f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$ in the backtracking line search must be interpreted carefully.
- By our convention that f is infinite outside its domain, the inequality implies that $x + t\Delta x \in \mathbf{dom}\ f$.

Backtracking line search (5/5)

- In a practical implementation, we first multiply t by β until $x + t\Delta x \in \textbf{dom } f$; then we start to check whether the inequality

$$f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$$

  holds.
- The parameter α is typically chosen between 0.01 and 0.3, meaning that we accept a decrease in f between 1% and 30% of the prediction based on the linear extrapolation.
- The parameter β is often chosen to be between 0.1 (which corresponds to a very crude search) and 0.8 (which corresponds to a less crude search).

Gradient descent method

- A natural choice for the search direction is the negative gradient $\Delta x = -\nabla f(x)$. The resulting algorithm is called the **gradient algorithm**, **gradient method**, or **gradient descent method**.

- **Algorithm 3.** Gradient descent method.
  given a starting point $x \in \mathbf{dom}\ f$.
  repeat
  1. $\Delta x := -\nabla f(x)$.
  2. *Line search*. Choose step size t via exact or backtracking line search.
  3. *Update*. $x := x + t\Delta x$.

  until stopping criterion is satisfied.

- The stopping criterion is usually of the form $||\nabla f(x)||_2 \leq \eta$, where η is small and positive.

Convergence analysis (1/2)

- In the following, we present a simple convergence analysis for the gradient method with exact line search and backtracking line search.
- The lighter notation $x^+ = x + t\Delta x$ is adopted in place of $x^{(k+1)} = x^{(k)} + t^{(k)}\Delta x^{(k)}$, where $\Delta x = -\nabla f(x)$.
- We assume f is strongly convex on S, so there are positive constants m and M such that $mI \preceq \nabla^2 f(x) \preceq MI$ for all $x \in S$.

Convergence analysis (2/2)

- Define the function $\tilde{f} : \mathbf{R} \to \mathbf{R}$ by

$$\tilde{f}(t) = f(x - t\nabla f(x)),$$

  i.e., f as a function of the step length t in the negative gradient direction.

- Assume $x - t\nabla f(x) \in S$. From the inequality

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{M}{2}||y - x||_2^2,$$

  with $y = x - t\nabla f(x)$, we obtain a quadratic upper bound on $\tilde{f}$:

$$\tilde{f}(t) \leq f(x) - t||\nabla f(x)||_2^2 + \frac{Mt^2}{2}||\nabla f(x)||_2^2.$$

Analysis for exact line search (1/2)

- For the case of exact line search, we will show that

$$f(x^{(k)}) - p^* \leq c^k (f(x^{(0)}) - p^*)$$

where $c = 1 - m/M < 1$.

- Minimizing over t both sides of the inequality

$$\tilde{f}(t) \leq f(x) - t||\nabla f(x)||_2^2 + \frac{Mt^2}{2}||\nabla f(x)||_2^2,$$

we get

$$f(x^+) = \tilde{f}(t_{\text{exact}}) \leq f(x) - \frac{1}{2M}||\nabla(f(x))||_2^2.$$

Analysis for exact line search (2/2)

- Subtracting p^* from both sides, we get

$$f(x^+) - p^* \leq f(x) - p^* - \frac{1}{2M}||\nabla f(x)||_2^2.$$

- Recall that $||\nabla f(x)||_2^2 \geq 2m(f(x) - p^*)$. So

$$f(x^+) - p^* \leq (1 - m/M)(f(x) - p^*).$$

Analysis for backtracking line search (1/4)

- For backtracking line search, we will show that

$$f(x^{(k)}) - p^* \leq c^k (f(x^{(0)}) - p^*)$$

where

$$c = 1 - \min\{2m\alpha, 2\beta\alpha m/M\} < 1.$$

- We first show that the backtracking exit condition,

$$\tilde{f}(t) \leq f(x) - \alpha t ||\nabla f(x)||_2^2,$$

is satisfied whenever $0 \leq t \leq 1/M$.

Analysis for backtracking line search (2/4)

- To see this, first note that $0 \leq t \leq 1/M$ implies $-t + \frac{Mt^2}{2} \leq -t/2$.
- Then, starting from a previously derived bound for $\tilde{f}(t)$,

$$\tilde{f}(t) \leq f(x) - t||\nabla f(x)||_2^2 + \frac{Mt^2}{2}||\nabla(f(x))||_2^2,$$

we can see that, for $0 \leq t \leq 1/M$,

$$\begin{aligned} \tilde{f}(t) &\leq f(x) + \left(-t + \frac{Mt^2}{2}\right)||\nabla f(x)||_2^2, \\ &\leq f(x) - (t/2)||\nabla f(x)||_2^2 \\ &\leq f(x) - \alpha t||\nabla f(x)||_2^2. \end{aligned}$$

Analysis for backtracking line search (3/4)

- Therefore the backtracking line search terminates either with $t = 1$ or with a value $t \geq \beta/M$.
- In the first case we have

$$f(x^+) \leq f(x) - \alpha||\nabla f(x)||_2^2,$$

and in the second case we have

$$f(x^+) \leq f(x) - (\beta\alpha/M)||\nabla f(x)||_2^2.$$

- Putting these together, we always have

$$f(x^+) \leq f(x) - \min\{\alpha, \beta\alpha/M\}||\nabla f(x)||_2^2.$$

Analysis for backtracking line search (4/4)

- From

$$f(x^+) \leq f(x) - \min\{\alpha, \beta\alpha/M\}||\nabla f(x)||_2^2,$$

we can follow a similar derivation and conclude that

$$f(x^{(k)}) - p^* \leq c^k(f(x^{(0)}) - p^*)$$

where

$$c = 1 - \min\{2m\alpha, 2\beta\alpha m/M\} < 1.$$

Example – A quadratic problem in $\mathbf{R}^2$ (1/4)

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

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

where $\gamma > 0$.

- Clearly, the optimal point is $x^* = 0$, and the optimal value is 0.
- The Hessian of f is constant, and has eigenvalues 1 and γ, so the condition numbers of the sublevel sets of f are all exactly

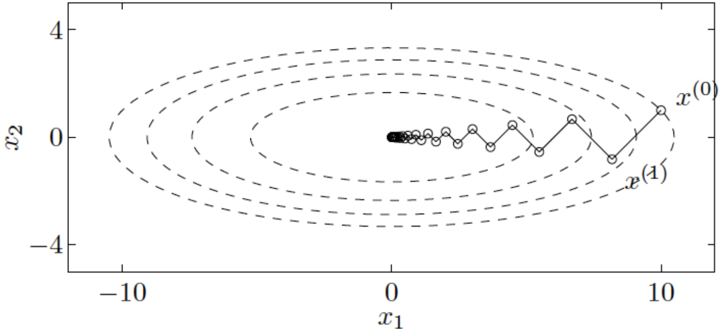
$$\frac{\max\{1, \gamma\}}{\min\{1, \gamma\}} = \max\{\gamma, 1/\gamma\}.$$

Example – A quadratic problem in $\mathbf{R}^2$ (2/4)

- The tightest choices for the strong convexity constants m and M are
$$m = \min\{1, \gamma\}, M = \max\{1, \gamma\}.$$
- We apply the gradient descent method with exact line search, starting at the point $x^{(0)} = (\gamma, 1)$.
- It can be shown that the kth iterate $x^{(k)}$ has the closed-form expression as follows:
$$x_1^{(k)} = \gamma\left(\frac{\gamma-1}{\gamma+1}\right)^k, \quad x_2^{(k)} = \left(-\frac{\gamma-1}{\gamma+1}\right)^k,$$
and the corresponding function value is
$$f(x^{(k)}) = \frac{\gamma(\gamma+1)}{2}\left(\frac{\gamma-1}{\gamma+1}\right)^{2k} = \left(\frac{\gamma-1}{\gamma+1}\right)^{2k} f(x^{(0)}).$$

Example – A quadratic problem in $\mathbf{R}^2$ (3/4)

- This case for $\gamma = 10$ is illustrated below.



Example – A quadratic problem in $\mathbf{R}^2$ (4/4)

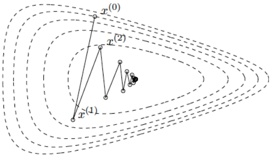
- For this simple example, convergence is exactly linear, i.e., the error is exactly a geometric series, reduced by the factor $|(\gamma - 1)/(\gamma + 1)|^2$ at each iteration.
- For $\gamma = 1$, the exact solution is found in one iteration; for γ not far from one (say, between $1/3$ and 3) convergence is rapid.
- The convergence is very slow for $\gamma \gg 1$ or $\gamma \ll 1$.

Example – A nonquadratic problem in $\mathbf{R}^2$ (1/6)

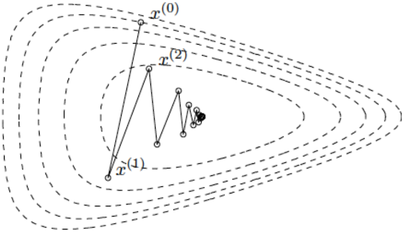
- We now consider a nonquadratic example in $\mathbf{R}^2$, with

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

- We apply the gradient method with a backtracking line search, with $\alpha = 0.1, \beta = 0.7$.
- The following figure shows some level curves of f, and the iterates $x^{(k)}$ generated by the gradient method (shown as small circles).



Example – A nonquadratic problem in $\mathbf{R}^2$ (2/6)

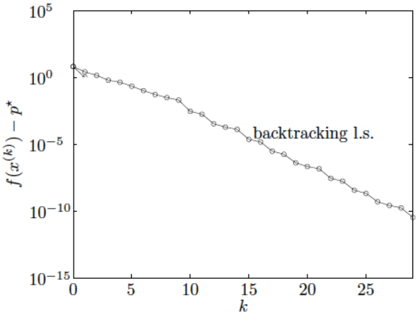


- The lines connecting successive iterates show the scaled steps,

$$x^{(k+1)} - x^{(k)} = -t^{(k)} \nabla f(x^{(k)}).$$

Example – A nonquadratic problem in $\mathbf{R}^2$ (3/6)

- The figure below shows the error $f(x^{(k)}) - p^*$ versus iteration k.

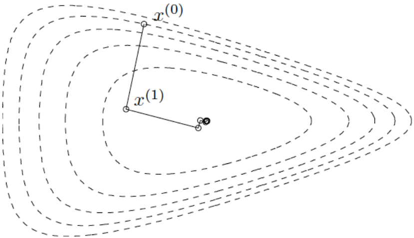


Example – A nonquadratic problem in $\mathbf{R}^2$ (4/6)

- The error converges to zero approximately as a geometric series.
- In 20 iterations, the error is reduced from about 10 to about 10^{-7}, so the error is reduced by a factor of approximately $10^{-8/20} \approx 0.4$ each iteration.
- This reasonably rapid convergence is predicted by our convergence analysis, since the sublevel sets of f are not too badly conditioned, which in turn means that M/m can be chosen as not too large.

Example – A nonquadratic problem in $\mathbf{R}^2$ (5/6)

- To compare backtracking line search with an exact line search, we use the gradient method with an exact line search, on the same problem, and with the same starting point.
- The results are given in the following figure. Here too the convergence is approximately linear, about twice as fast as the gradient method with backtracking line search.



Example – A nonquadratic problem in $\mathbf{R}^2$ (6/6)

- With exact line search, the error is reduced by about 10^{-11} in 15 iterations, i.e., a reduction by a factor of about $10^{-11/15} \approx 0.2$ per iteration.