

**Solutions 2**

Fall 2009

**Reading:** Sections 1.2–1.3 of Nonlinear programming by Bertsekas. (**Note:** Be sure to use the scanned PDFs on webpage to obtain the correct numbering of the problems in this assignment.)

**Solution 2.1**

The problem is to perform an unconstrained minimization of the function  $f(x, y) = 3x^2 + y^4$  using various methods. Using the starting point  $x^0 = (1, -2)$ , we have

$$f(x^0) = 19, \quad \nabla f(x^0) = [6x \quad 4y^3] = [6 \quad -32]$$

(a) Steepest descent with  $s = 1$ ,  $\sigma = 0.1$  and  $\beta = 0.5$  takes the form:

$$x^1 = x^0 + \beta^m s d^0$$

where  $d^0 = -\nabla f(x^0)$ . Here are the results of determining the appropriate  $m$  according to the Armijo rule:

$m$	$\beta^m s$	$f(x^0) - f(x^0 + \beta^m s d^0)$	$-\sigma \beta^m s \nabla f(x^0)^T d^0$
0	1	-810056	106
1	.5	-38409	53
2	.25	-1277.75	26.5
3	.125	2.8125	13.25
4	.0625	17.828125	6.625

So one iteration of steepest descent requires 5 internal iterations to determine the value of  $m$ ; it yields the new point  $x^1 = (0.625, 0)$  where  $f(x^1) = 1.17$ .

(b) Steepest descent with  $s = 1$ ,  $\sigma = 0.1$  and  $\beta = 0.1$ .

$m$	$\beta^m s$	$f(x^0) - f(x^0 + \beta^m s d^0)$	$-\sigma \beta^m s \nabla f(x^0)^T d^0$
0	1	-810056	106
1	.1	-16.4464	10.6

Now only 2 internal iterations are required, and yield the new point  $x^1 = (0.4, 1.2)$  where  $f(x^1) = 2.55$ . The smaller value of  $\beta$  reduced the amount of time to find the stepsize, but yielded a new point  $x^1$  with higher cost.

(c) Newton's method with  $s = 1$ ,  $\sigma = 0.1$  and  $\beta = 0.5$ . We now use the descent direction  $d^0 = -(\nabla^2 f(x^0))^{-1} \nabla f(x^0)$ . Calculations yield

$$d^0 = - \begin{bmatrix} 6 & 0 \\ 0 & 48 \end{bmatrix}^{-1} [6 \quad 32] = [-1 \quad 2/3].$$

$m$	$\beta^m s$	$f(x^0) - f(x^0 + \beta^m s d^0)$	$-\sigma \beta^m s \nabla f(x^0)^T d^0$
0	1	15.8395	2.733

So one iteration of Newton's method requires only 1 internal iteration to determine  $m$ , and yields  $x^1 = (0, -4/3)'$  with  $f(x^1) = 3.1605$ . Although fewer internal iterations were required, the cost of the resulting  $x^1$  is higher. In this case, determining the descent direction (involving the inverse of the Hessian) was straightforward, but this calculation is potentially expensive.

### Solution 2.2

Since  $\nabla f(x) = (2 + \beta)\|x\|^\beta x$ , we have

$$x^{k+1} = (1 - s(2 + \beta)\|x^k\|^\beta)x^k \Rightarrow \|x^{k+1}\| = \left|1 - s(2 + \beta)\|x^k\|^\beta\right| \|x^k\|$$

Let  $\gamma^k = 1 - s(2 + \beta)\|x^k\|^\beta$ . So  $\gamma^k \leq 1, \forall k$  and we also have  $\|x^{k+1}\| = |\gamma^k| \|x^k\|$

- If  $\gamma^0 < -1$  then  $\|x^1\| > \|x^0\|$ , so  $\gamma^1 < \gamma^0 < -1$ . Similarly,  $\gamma^k < -1, \forall k$ . Therefore  $\{\|x^k\|\}$  is strictly increasing, hence it does not converge.
- If  $\gamma^0 = -1$  then  $x^1 = -x^0$ , similarly  $x^{k+1} = -x^k$ . So  $\{x^k\}$  converges only if  $x^0 = 0$  which contradicts with  $\gamma^0 = -1$
- If  $-1 < \gamma^0 < 1$  ( $s > 0$ ), then  $\|x^1\| < \|x^0\|$ , so  $\gamma^1 > \gamma^0 > -1$ . Similarly,  $\gamma^k > -1, \forall k$ . Therefore  $\|x^k\|$  is decreasing and bounded below (by 0), so it converges to  $x^*$ .

$$x^* = (1 - s(2 + \beta)\|x^*\|^\beta)x^* \Rightarrow x^* = 0$$

- If  $\gamma^0 = 1$  then  $x^1 = x^0$ , thus  $x^k = x^0, \forall k$ . The sequence also converges to  $x^0$

In summary,  $\{x^k\}$  converges to  $x^* = 0$  if and only if  $x_0 = 0$  or  $0 < s < \frac{2}{(2 + \beta)\|x^0\|^\beta}$

### Solution 2.3

Notice  $\nabla f(x) = \frac{3}{2}\|x\|^{-\frac{1}{2}}x$ , and let  $y = -x$  then

$$\frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|} = \frac{3}{2}\|x\|^{-\frac{1}{2}} \rightarrow \infty \text{ as } x \rightarrow 0$$

Therefore, the Lipschitz condition does not hold.

It is easy to see that if  $\|x^k\| = \frac{9s^2}{4}$  for some  $k$  then the steepest descent algorithm converges to 0 in finite iterations.

It remains to show that the algorithm does not converges in infinite steps. If this is not true, then

$$\forall \epsilon > 0, \exists N = N(\epsilon) : \forall k \geq N, \|x^k\| < \epsilon$$

Let  $\epsilon = \left(\frac{s}{2}\right)^2$ , then

$$\|x^{k+1}\| = \left|1 - \frac{3}{2}s\|x^k\|^{-1/2}\right| \|x^k\| > 2\|x^k\|$$

So eventually,  $\|x^k\| > \epsilon$  which is a contradiction.

**Solution 2.4**

Our approach is to show that  $\{d^k\}$  thus defined satisfies the gradient-relatedness condition. Let  $\{x^k\}$  be any subsequence that converges to a non-stationary point  $\bar{x}$  (so that  $\nabla f(\bar{x}) \neq 0$ ). Since  $f$  is continuously differentiable, we also have  $\nabla f(x^k) \rightarrow \nabla f(\bar{x})$ . Now the definition of the direction  $d^k$  yields that

$$\|d_k\|_2^2 = \sum_{i=1}^n (d_i^k)^2 = [\|\nabla f(x^k)\|_\infty]^2$$

where  $\|\nabla f(x^k)\|_\infty := \max_{i=1, \dots, n} |\frac{\partial f}{\partial x_i}(x^k)|$ . Since  $\nabla f(x^k)$  converges to  $\nabla f(\bar{x})$ , it is bounded and hence  $d^k$  is also bounded. Moreover, it holds that  $\nabla f(x^k)^T d^k = -\|\nabla f(x^k)\|_\infty^2$ . Thus,

$$\begin{aligned} \frac{\nabla f(x^k)^T d^k}{\|\nabla f(x^k)\|_2 \|d^k\|_2} &= \frac{-\|\nabla f(x^k)\|_\infty^2}{\|\nabla f(x^k)\|_2 \|\nabla f(x^k)\|_\infty} \\ &= \frac{-\|\nabla f(x^k)\|_\infty}{\|\nabla f(x^k)\|_2}, \end{aligned}$$

which stays bounded away from zero since  $\nabla f(x^k)$  converges to a non-stationary point by assumption. Thus, the sequence  $\{d^k\}$  is gradient-related, so that every limit point must be stationary by the result proved in class.

**Solution 2.5**

Notice that  $\nabla f(x) = Qx$ , so

$$x^{k+1} = x^k - \alpha Qx^k = (I - \alpha Q)x^k \Rightarrow x^k = (I - \alpha Q)^k x^0$$

Note that if  $(\lambda, u)$  is eigen-pair of  $Q$  then  $(1 - \alpha\lambda, u)$  is an eigen-pair of  $A := I - \alpha Q$ . Since  $Q$  is invertible, its eigenvectors form a basis of  $R^n$ , so  $x^0$  can be written as  $x^0 = \sum_{i=1}^n \beta_i u_i$ , so

$$x^k = A^k x^0 = \sum_{i=1}^n \beta_i A^k u_i = \sum_{i=1}^n \beta_i (1 - \alpha\lambda_i)^k u_i$$

Notice that if  $\lambda_i < 0$  then  $1 - \alpha\lambda_i > 1$  so  $(1 - \alpha\lambda_i)^k \rightarrow \infty$  as  $k \rightarrow \infty$ .

Therefore, unless  $x^0$  belongs to the subspace spanned by the eigenvectors of  $Q$  corresponding to the non-negative eigenvalues, the sequence  $\{x^k\}$  diverges.

**Solution 2.6**

$$f(x, y) = \frac{1}{2} \begin{pmatrix} x \\ y \end{pmatrix}^T Q \begin{pmatrix} x \\ y \end{pmatrix}$$

where  $Q = \begin{pmatrix} 2 & 1.999 \\ 1.999 & 2 \end{pmatrix}$ . Eigenvalues of  $Q$  are  $m = 0.001$  and  $M = 3.999$

The rate of convergence is

$$\left( \frac{M - m}{M + m} \right)^2 = 0.9990$$

A starting point for which the convergence rate is sharp:

$$x^0 = U \begin{pmatrix} 1/m \\ 1/M \end{pmatrix} = \begin{pmatrix} -706.9300 \\ 707.2836 \end{pmatrix}$$

(By transforming  $u = Ux$  where  $Q = U\Lambda U^T$  is the eigen-decomposition of  $Q$ , we get  $f(x) = g(u) = mu_1^2 + Mu_2^2$ )

(See Figure 1.3.2, p61 *Nonlinear programming*)

### Solution 2.7

Without loss of generality, we assume that  $x^* = 0$ . (If not, we simply apply the transformation  $y = x - x^*$  to the problem.) The update then takes the form

$$x^{k+1} = (I - sQ)x^k - se^k.$$

Using the triangle inequality, we have

$$\begin{aligned} \|x^k\| &\leq \|(I - sQ)x^{k-1}\| + s\|e^{k-1}\| \\ &\leq q\|x^{k-1}\| + s\delta \end{aligned}$$

where  $\|(I - sQ)x^{k-1}\| \leq q\|x^{k-1}\|$  follows from the same argument as used in class. Applying this inequality recursively yields

$$\begin{aligned} \|x^k\| &\leq q \left[ q\|x^{k-2}\| + s\delta \right] + s\delta \\ &\vdots \\ &\leq q^k \|x^0\| + s\delta \sum_{i=0}^{k-1} q^i \\ &\leq q^k \|x^0\| + \frac{s\delta}{1-q} \end{aligned}$$

where the final line uses the fact that for  $q \in [0, 1)$

$$\sum_{i=0}^{k-1} q^i \leq \sum_{i=0}^{\infty} q^i = \frac{1}{1-q}.$$

### Solution 2.8

(a) Note that  $x^* = -Q^{-1}c$  is the optimal solution. So by a change of variables (consider  $x \leftarrow x - x^*$ ), we can assume  $c = 0$  WLOG.

Since  $\nabla f(x) = Qx$ , we have

$$x^{k+1} = x^k - \alpha Qx^k + \beta(x^k - x^{k-1}) = ((1 + \beta)I - \alpha Q)x^k - \beta x^{k-1}$$

So

$$\begin{pmatrix} x^{k+1} \\ x^k \end{pmatrix} = \begin{pmatrix} (1 + \beta)I - \alpha Q & -\beta I \\ I & 0 \end{pmatrix} \begin{pmatrix} x^k \\ x^{k-1} \end{pmatrix}$$

Let  $y^k = \begin{pmatrix} x^k \\ x^{k-1} \end{pmatrix}$  and  $A = \begin{pmatrix} (1+\beta)I - \alpha Q & -\beta I \\ I & 0 \end{pmatrix}$ , above equation is equivalent to

$$y^{k+1} = Ay^k$$

Suppose  $v$  is eigenvalue of  $A$ ,

$$\begin{pmatrix} (1+\beta)I - \alpha Q & -\beta I \\ I & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = v \begin{pmatrix} x \\ y \end{pmatrix} \Rightarrow \alpha Qx = (1+\beta - v - \beta/v)x$$

Hence  $1 + \beta - v - \beta/v = \alpha\lambda$  (\*) for some eigenvalue  $\lambda$  of  $Q$ .

If  $0 < \alpha < \frac{2(1+\beta)}{M} \leq \frac{2(1+\beta)}{\lambda_i(Q)}$  then  $0 < \alpha\lambda_i(Q) < 2(1+\beta), \forall i$ .

Therefore,  $-(1+\beta) < \lambda_i(A) + \beta/\lambda_i(A) < (1+\beta)$ , which implies  $|\lambda_i(A)| < 1, \forall i$ .

Let  $\gamma = \|A\|_2 = \max\{|\lambda_i(A)|\} < 1$ , we have

$$\|y^{k+1}\| = \|Ay^k\| \leq \|A\|_2 \|y^k\| = \gamma \|y^k\|$$

The sequence  $\{y^k\}$  thus converges linearly with the rate of  $\gamma$ , so is  $\{x^k\}$ .

By solving (\*) for  $v$ , we have

$$v = \frac{1}{2} \left( 1 + \beta - \alpha\lambda \pm \sqrt{(1 + \beta - \alpha\lambda)^2 - 4\beta} \right)$$

(Note that  $v$  might be complex number.)

Therefore,

$$\gamma = \max_{\lambda \in \{\lambda_i(Q)\}} \frac{1}{2} \left| 1 + \beta - \alpha\lambda \pm \sqrt{(1 + \beta - \alpha\lambda)^2 - 4\beta} \right|$$

We choose  $\alpha, \beta$  as following

$$\alpha = \frac{4}{(\sqrt{M} + \sqrt{m})^2}, \beta = \left( \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \right)^2$$

We have, for any  $\lambda \in \{\lambda_i(Q)\}$

$$\begin{aligned} & \frac{1}{2} \left| 1 + \beta - \alpha\lambda \pm \sqrt{(1 + \beta - \alpha\lambda)^2 - 4\beta} \right| \\ &= \frac{1}{2} \left| \frac{2(M+m-2\lambda)}{(\sqrt{M} + \sqrt{m})^2} \pm \sqrt{\frac{4(M+m-2\lambda)^2}{(\sqrt{M} + \sqrt{m})^4} - 4 \left( \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \right)^2} \right| \\ &= \frac{1}{(\sqrt{M} + \sqrt{m})^2} \left| (M+m-2\lambda) \pm \sqrt{(M+m-2\lambda)^2 - (M-m)^2} \right| \\ &= \frac{1}{(\sqrt{M} + \sqrt{m})^2} \left| (M+m-2\lambda) \pm 2\sqrt{(M-\lambda)(m-\lambda)} \right| \\ &= \frac{1}{(\sqrt{M} + \sqrt{m})^2} \left| (M+m-2\lambda) \pm 2i\sqrt{(M-\lambda)(\lambda-m)} \right| \quad (\text{since } m \leq \lambda \leq M) \\ &= \frac{1}{(\sqrt{M} + \sqrt{m})^2} \sqrt{(M+m-2\lambda)^2 + 4(M-\lambda)(\lambda-m)} \\ &= \frac{M-m}{(\sqrt{M} + \sqrt{m})^2} = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}} \end{aligned}$$

Therefore the corresponding convergence rate is

$$\gamma = \frac{\sqrt{M} - \sqrt{m}}{\sqrt{M} + \sqrt{m}}$$

**Note:** To find optimal  $\alpha, \beta$  we would solve the following problem

$$(\alpha^*, \beta^*) = \arg \min_{\alpha, \beta} \max_{\lambda \in \{\lambda_i(Q)\}} \frac{1}{2} \left| 1 + \beta - \alpha\lambda \pm \sqrt{(1 + \beta - \alpha\lambda)^2 - 4\beta} \right|$$

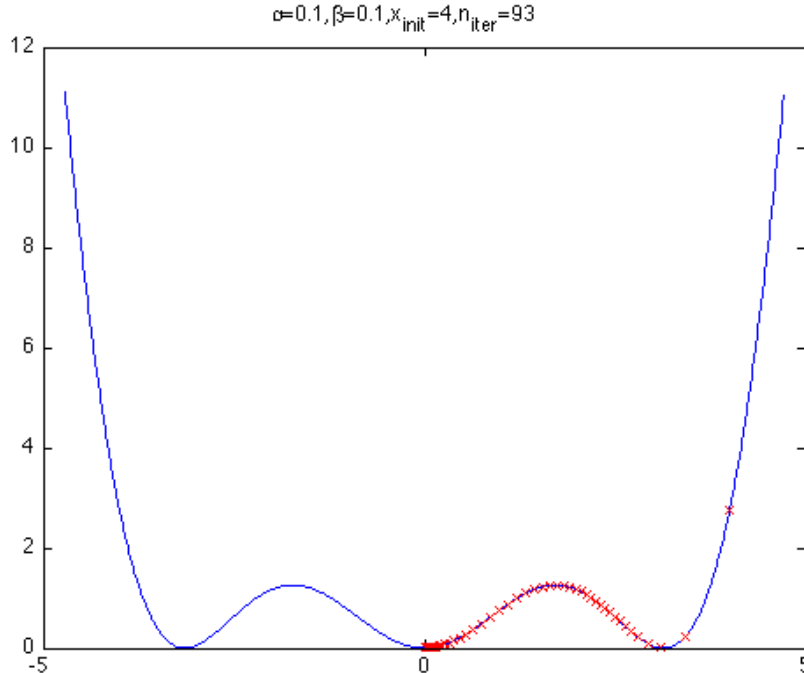
With careful derivations (complex number issue), one can prove that the above values for  $\alpha, \beta$  are the optimal solution.

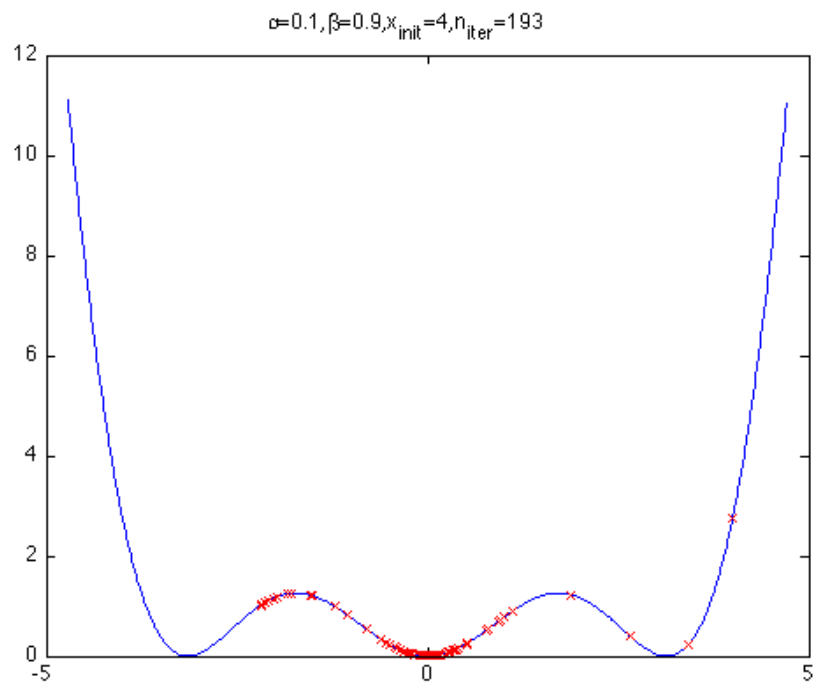
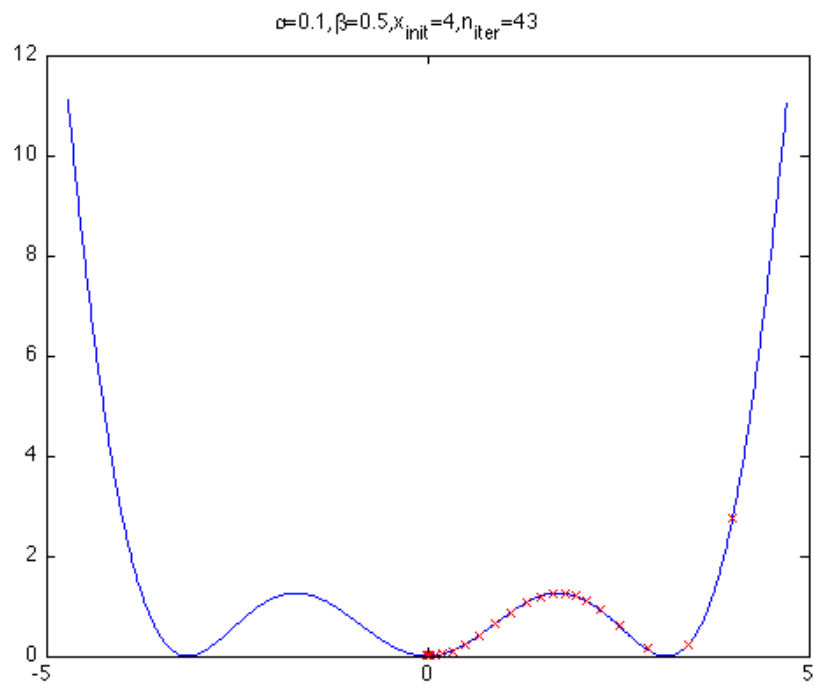
In many cases, the solution of  $\min_{\theta} \max_{\lambda \in \{\lambda_i\}_{i=1}^n} f(\theta, \lambda)$  often satisfies  $f(\theta^*, \lambda_i) = c, \forall i$ .

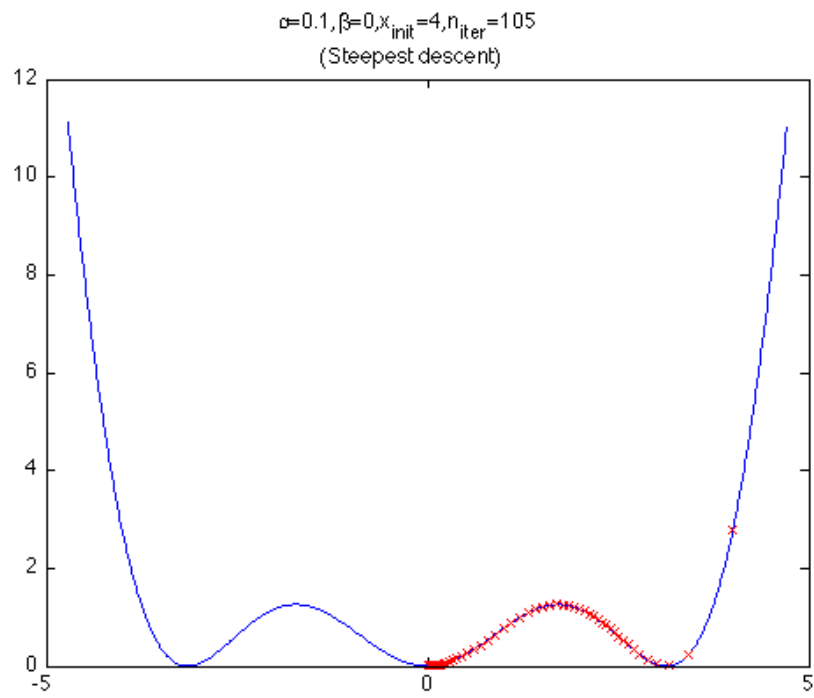
- (b) Assuming that the parameter  $\beta$  is appropriately chosen, it is a reasonable conjecture that the method would have better behavior for functions with varying steepness. We provide empirical support for this conjecture with our experimental results reported in the following part.
- (c) For  $f(x) = \frac{1}{2}x^2(1 + \gamma \cos(x))$ ,

$$f'(x) = x + \gamma x \cos(x) - \frac{1}{2}\gamma x^2 \sin(x).$$

The following four plots show successive iterates and the total number of iterations for four different choices of  $\beta$ , where the final plot shows  $\beta = 0$ , corresponding to steepest descent. For the intermediate value of  $\beta = 0.5$ , fewer iterations are required overall. For too large values of  $\beta$ , the method overshoots.

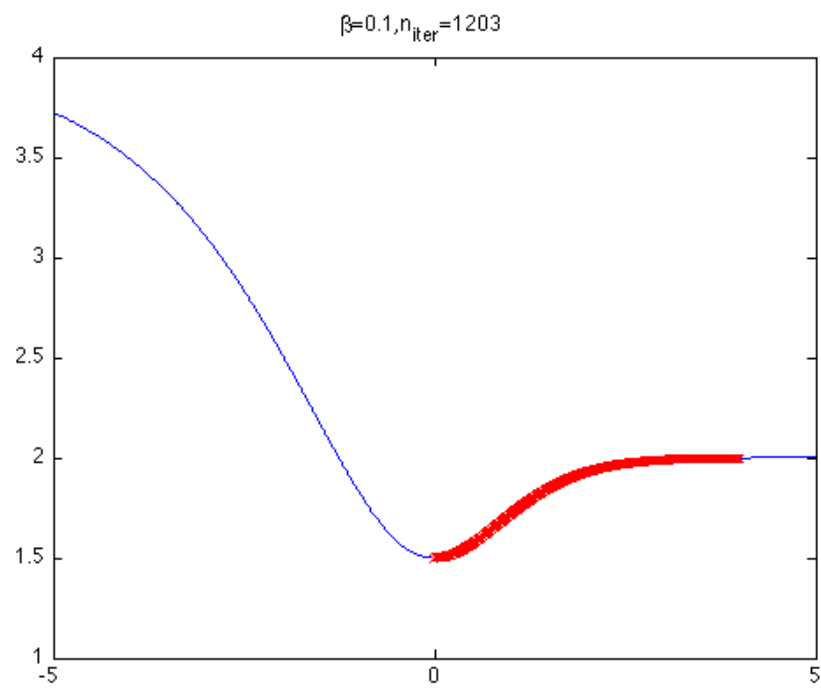


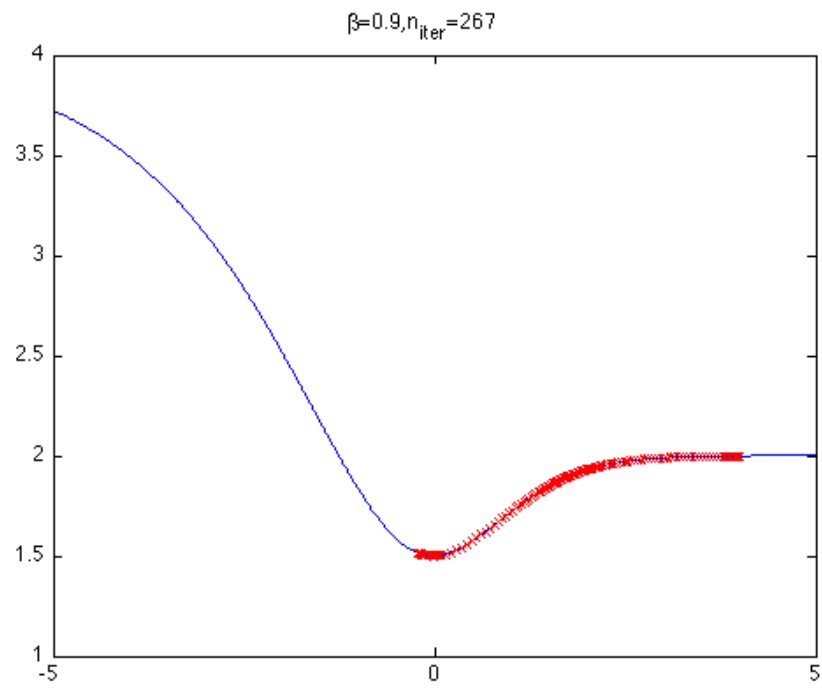
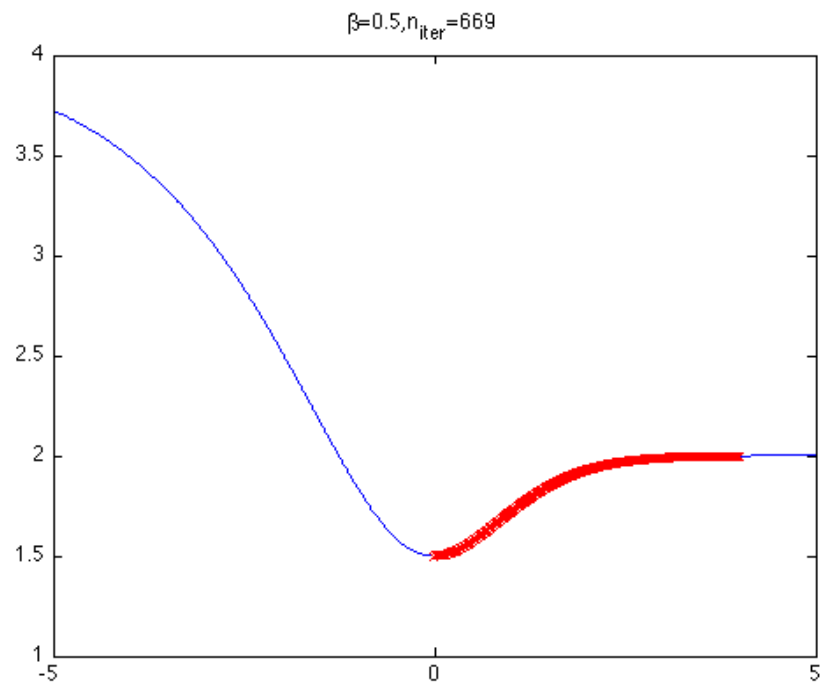


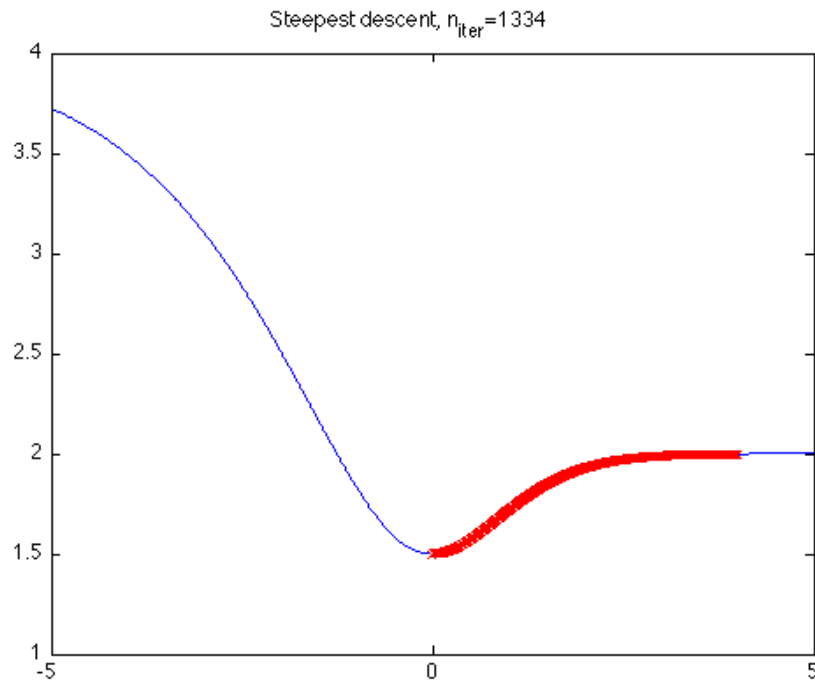


For the function  $f(x) = \frac{1}{2} \sum_{i=1}^m |z_i - \tanh(xy_i)|^2$ , we have

$$f'(x) = \sum_{i=1}^n (\tanh(xy_i) - z_i)(1 - \tanh(xy_i)^2)y_i$$







```

function [x_opt, n_iter] = heavy_ball(alpha, beta, epsilon, x_init)
    % Replace your function definition here
    gamma = 0.5;
    f = @(x) 0.5*x.^2.*(1+cos(x));
    grad_f = @(x) x + gamma*x*(cos(x) - 0.5*x*sin(x));

    range = [-1.5*pi:0.01:1.5*pi];
    plot(range, f(range)); hold on;

    x = [x_init x_init];
    dist = inf;
    n_iter = 0;
    while dist > epsilon
        x_new = x(end) - alpha*grad_f(x(end)) + beta*(x(end)-x(end-1));
        dist = abs(x_new - x(end));
        x = [x x_new];
        n_iter = n_iter + 1;
    end
    x_opt = x(end);
    plot(x, f(x), 'rx'); hold on;
end

```