General Algorithm: Steepest Descent (contd)

Find a starting point $\mathbf{x}^{(0)} \in \mathcal{D}$.
repeat

1. Set $\Delta\mathbf{x}^{(k)} = \text{argmin}\left\{\nabla^T f(\mathbf{x}^{(k)})\mathbf{v} \mid ||\mathbf{v}|| = 1\right\}$.
2. Choose a step size $t^{(k)} > 0$ using exact or backtracking ray search.
3. Obtain $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\Delta\mathbf{x}^{(k)}$.
4. Set $k = k + 1$.

until stopping criterion (such as $||\nabla f(\mathbf{x}^{(k+1)})|| \leq \epsilon$) is satisfied

Figure 9: The steepest descent algorithm.

Two examples of the steepest descent method are the gradient descent method (for the eucledian or L_2 norm) and the coordinate-descent method (for the L_1 norm). One fact however is that no two norms should give exactly opposite steepest descent directions, though they may point in different directions.

Algorithms: Coordinate-Descent Method

- Corresponds exactly to the choice of L_1 norm for the steepest descent method. The steepest descent direction using the L_1 norm is given by $\Delta\mathbf{x} = -\frac{\partial f(\mathbf{x})}{\partial x_i}\mathbf{u}^i$ where, $\frac{\partial f(\mathbf{x})}{\partial x_i} = ||\nabla f(\mathbf{x})||_\infty$ and $\mathbf{u}^i$ is defined as the unit vector pointing along the i^{th} axis.
- Thus each iteration of the coordinate descent method involves optimizing over one component of the vector $\mathbf{x}^{(k)}$ (having the largest absolute value in the gradient vector).

Find a starting point $\mathbf{x}^{(0)} \in \mathcal{D}$.
Select an appropriate norm $||.||$.
repeat

1. Let $\frac{\partial f(\mathbf{x}^{(k)})}{\partial x_i^{(k)}} = ||\nabla f(\mathbf{x}||_\infty)$.
2. Set $\Delta\mathbf{x}^{(k)} = -\frac{\partial f(\mathbf{x}^{(k)})}{\partial x_i^{(k)}}\mathbf{u}^i$.
3. Choose a step size $t^{(k)} > 0$ using exact or backtracking ray search.
4. Obtain $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\Delta\mathbf{x}^{(k)}$.
5. Set $k = k + 1$.

until stopping criterion (such as $||\nabla f(\mathbf{x}^{(k+1)})||_\infty \leq \epsilon$) is satisfied

Algorithms: Gradient Descent

- This classic greedy algorithm for minimization uses the negative of the gradient of the function at the current point $\mathbf{x}^*$ as the descent direction $\Delta\mathbf{x}^*$.
- This choice of $\Delta\mathbf{x}^*$ corresponds to the direction of steepest descent under the L_2 (eucledian) norm and follows from

Algorithms: Gradient Descent

- This classic greedy algorithm for minimization uses the negative of the gradient of the function at the current point $\mathbf{x}^*$ as the descent direction $\Delta\mathbf{x}^*$.
- This choice of $\Delta\mathbf{x}^*$ corresponds to the direction of steepest descent under the L_2 (eucledian) norm and follows from the Cauchy Shwarz inequality

Find a starting point $\mathbf{x}^{(0)} \in \mathcal{D}$
repeat
1. Set $\Delta\mathbf{x}^{(k)} = -\nabla f(\mathbf{x}^{(k)})$.
2. Choose a step size $t^{(k)} > 0$ using exact or backtracking ray search.
3. Obtain $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\Delta\mathbf{x}^{(k)}$.
4. Set $k = k + 1$.

until stopping criterion (such as $||\nabla f(\mathbf{x}^{(k+1)})||_2 \leq \epsilon$) is satisfied

The steepest descent method can be thought of as changing the coordinate system in a particular way and then applying the gradient descent method in the changed coordinate system.

Convergence of the Gradient Descent Algorithm

- We recap the (necessary) inequality (34) resulting from Lipschitz continuity of $\nabla f(\mathbf{x})$. $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2$
- Considering $\mathbf{x}^k \equiv \mathbf{x}$, and $\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \nabla f(\mathbf{x}^k) \equiv \mathbf{y}$, we get

Convergence of the Gradient Descent Algorithm

- We recap the (necessary) inequality (34) resulting from Lipschitz continuity of $\nabla f(\mathbf{x})$.$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2$
- Considering $\mathbf{x}^k \equiv \mathbf{x}$, and $\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \nabla f(\mathbf{x}^k) \equiv \mathbf{y}$, we get

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - t^k \nabla^\top f(\mathbf{x}^k) \nabla f(\mathbf{x}^k) + \frac{L\left(t^k\right)^2}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$$

$$\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - (1 - \frac{Lt^k}{2})t\left\|\nabla f(\mathbf{x}^k)\right\|^2$$

t

- We have (44) if....

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{\widehat{t}}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 \quad (44)$$

if t^k <= 1/L and hat{t} = t (fixed step)
or t^k = min{....} and hat{t} = min value

Convergence of the Gradient Descent Algorithm

- We recap the (necessary) inequality (34) resulting from Lipschitz continuity of $\nabla f(\mathbf{x})$.$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2$
- Considering $\mathbf{x}^k \equiv \mathbf{x}$, and $\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \nabla f(\mathbf{x}^k) \equiv \mathbf{y}$, we get

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - t^k \nabla^\top f(\mathbf{x}^k) \nabla f(\mathbf{x}^k) + \frac{L\left(t^k\right)^2}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$$

$$\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - (1 - \frac{Lt^k}{2})t\left\|\nabla f(\mathbf{x}^k)\right\|^2$$

- We have (44) if....

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{\widehat{t}}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 \tag{44}$$

  - With fixed step size $t^k = \widehat{t}$, we ensure that $0 < \widehat{t} \leq \frac{1}{L}$ Question: Does it require me to know L?

Convergence of the Gradient Descent Algorithm

- We recap the (necessary) inequality (34) resulting from Lipschitz continuity of $\nabla f(\mathbf{x})$.$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2$
- Considering $\mathbf{x}^k \equiv \mathbf{x}$, and $\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \nabla f(\mathbf{x}^k) \equiv \mathbf{y}$, we get

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - t^k \nabla^\top f(\mathbf{x}^k) \nabla f(\mathbf{x}^k) + \frac{L\left(t^k\right)^2}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$$

$$\Longrightarrow f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - (1 - \frac{Lt^k}{2}) t \left\|\nabla f(\mathbf{x}^k)\right\|^2$$

- We have (44) if....

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{\widehat{t}}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 \quad (44)$$

  - With fixed step size $t^k = \widehat{t}$, we ensure that $0 < \widehat{t} \leq \frac{1}{L} \Longrightarrow 1 - \frac{L\widehat{t}}{2} \geq \frac{1}{2}$.

Convergence of the Gradient Descent Algorithm

- We recap the (necessary) inequality (34) resulting from Lipschitz continuity of $\nabla f(\mathbf{x})$.$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2$
- Considering $\mathbf{x}^k \equiv \mathbf{x}$, and $\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \nabla f(\mathbf{x}^k) \equiv \mathbf{y}$, we get

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - t^k \nabla^\top f(\mathbf{x}^k) \nabla f(\mathbf{x}^k) + \frac{L\left(t^k\right)^2}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$$

$$\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - (1 - \frac{Lt^k}{2})t\left\|\nabla f(\mathbf{x}^k)\right\|^2$$

- We have (44) if....

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{\widehat{t}}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 \tag{44}$$

  - With fixed step size $t^k = \widehat{t}$, we ensure that $0 < \widehat{t} \leq \frac{1}{L} \implies 1 - \frac{L\widehat{t}}{2} \geq \frac{1}{2}$.
  - With backtracking step seach, (44) holds with $\widehat{t} = min\left\{1, \beta\frac{2(1-c_1)}{L}\right\}$

- Using convexity, we have $f(\mathbf{x}^*) \geq f(\mathbf{x}^k) + \nabla^\top f(\mathbf{x}^k)(\mathbf{x}^* - \mathbf{x}^k)$
  $\implies f(\mathbf{x}^k) \leq f(\mathbf{x}^*) + \nabla^\top f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*)$
- Thus,
  $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{t}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$
  $\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^*) + \nabla^\top f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) - \frac{t}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$
  $\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^*) + \frac{1}{2t}\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2 + \nabla^T f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) - \frac{t}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 - \frac{1}{2t}\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2$

||x^{k+1} - x^*||^2

- Using convexity, we have $f(\mathbf{x}^*) \geq f(\mathbf{x}^k) + \nabla^\top f(\mathbf{x}^k)(\mathbf{x}^* - \mathbf{x}^k)$
  $\implies f(\mathbf{x}^k) \leq f(\mathbf{x}^*) + \nabla^\top f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*)$
- Thus,
  $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{t}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$
  $\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^*) + \nabla^\top f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) - \frac{t}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$
  $\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^*) + \frac{1}{2t}\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2 + \nabla^T f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) - \frac{t}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 - \frac{1}{2t}\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2$
  $\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^*) + \frac{1}{2t}\left(\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2 - \left\|\mathbf{x}^k - \mathbf{x}^* - t\nabla f(\mathbf{x}^k)\right\|^2\right)$
  $\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^*) + \frac{1}{2t}\left(\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2 - \left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\|^2\right)$ inequality still holds with canonical t

$$\implies f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*) \leq \frac{1}{2t}\left(\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2 - \left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\|^2\right) \tag{45}$$

we want to characterize the change wrt to k explicitly
Hence we sum these inequalities until k

- Summing (45) over all iterations (since $-\left\|\mathbf{x}^{k+1}-\mathbf{x}^*\right\|^2 < 0$), we have

$$\sum_{i=1}\left(f(\mathbf{x}^i)-f(\mathbf{x}^*)\right) \leq \frac{1}{2t}\left(\left\|\mathbf{x}^{(0)}-\mathbf{x}^*\right\|^2\right))$$

- The ray[6] and line search ensure that $f(\mathbf{x}^{i+1}) \leq f(\mathbf{x}^i)$ $\forall i = 0, 1, \ldots, k$. We thus get

[6] By Armijo condition in (27), for some $0 < c_1 < 1$, $f(\mathbf{x}^{i+1}) \leq f(\mathbf{x}^i) + c_1 t^i \nabla^T f(\mathbf{x}^i)\Delta\mathbf{x}^i$

- Summing (45) over all iterations (since $-\left\|\mathbf{x}^{k+1}-\mathbf{x}^*\right\|^2 < 0$), we have

$$\sum_{i=1}\left(f(\mathbf{x}^i)-f(\mathbf{x}^*)\right) \le \frac{1}{2t}\left(\left\|\mathbf{x}^{(0)}-\mathbf{x}^*\right\|^2\right))$$

- The ray[6] and line search ensure that $f(\mathbf{x}^{i+1}) \le f(\mathbf{x}^i)\ \forall i = 0, 1, \ldots, k$. We thus get

$$f(\mathbf{x}^k)-f(\mathbf{x}^*) \le \frac{1}{k}\sum_{i=1}^{k}\left(f(\mathbf{x}^i)-f(\mathbf{x}^*)\right) \le \frac{\left\|\mathbf{x}^{(0)}-\mathbf{x}^*\right\|^2}{2tk}$$

- Thus, as $k\to\infty$, $f(\mathbf{x}^k)\to f(\mathbf{x}^*)$. This shows convergence for gradient descent.

To ensure that f(x^k) - f(x^*) <= epsilon, we need
k = O(1/epsilon) [Terrible rate/order of convergence..]

[6] By Armijo condition in (27), for some $0 < c_1 < 1$, $f(\mathbf{x}^{i+1}) \le f(\mathbf{x}^i) + c_1 t^i \nabla^T f(\mathbf{x}^i)\Delta\mathbf{x}^i$

- Summing (45) over all iterations (since $-\left\|\mathbf{x}^{k+1}-\mathbf{x}^*\right\|^2 < 0$), we have

$$\sum_{i=1}\left(f(\mathbf{x}^i)-f(\mathbf{x}^*)\right) \leq \frac{1}{2t}\left(\left\|\mathbf{x}^{(0)}-\mathbf{x}^*\right\|^2\right))$$

- The ray[6] and line search ensure that $f(\mathbf{x}^{i+1}) \leq f(\mathbf{x}^i)\ \forall i = 0, 1, \ldots, k$. We thus get

$$f(\mathbf{x}^k)-f(\mathbf{x}^*) \leq \frac{1}{k}\sum_{i=1}^{k}\left(f(\mathbf{x}^i)-f(\mathbf{x}^*)\right) \leq \frac{\left\|\mathbf{x}^{(0)}-\mathbf{x}^*\right\|^2}{2tk}$$

- Thus, as $k \to \infty$, $f(\mathbf{x}^k) \to f(\mathbf{x}^*)$. This shows convergence for gradient descent.
- What we are more interested in however, is the **rate of convergence** of the gradient descent algorithm.

[6] By Armijo condition in (27), for some $0 < c_1 < 1$, $f(\mathbf{x}^{i+1}) \leq f(\mathbf{x}^i) + c_1 t^i \nabla^T f(\mathbf{x}^i)\Delta\mathbf{x}^i$

Aside: Backtracking ray search and Lipschitz Continuity

- Recap the Backtracking ray search algorithm
  - Choose a $\beta \in (0, 1)$
  - Start with $t = 1$
  - While $f(\mathbf{x} + t\Delta\mathbf{x}) > f(\mathbf{x}) + c_1 t \nabla^T f(\mathbf{x}) \Delta\mathbf{x}$, do
    - Update $t \leftarrow \beta t$

Aside: Backtracking ray search and Lipschitz Continuity [some justification for the magical hat{t}]

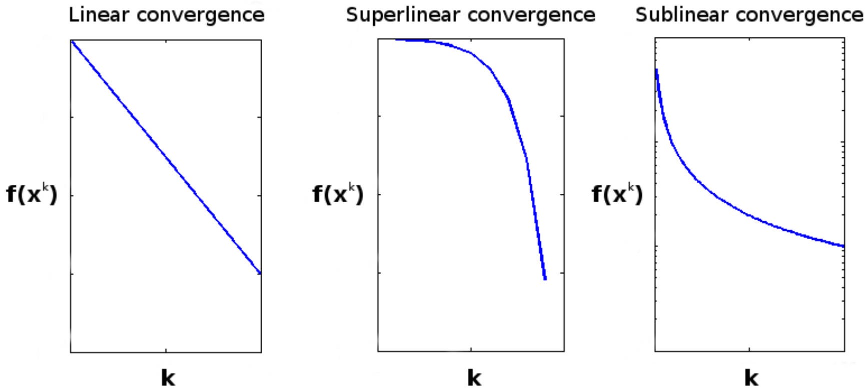
- Recap the Backtracking ray search algorithm
  - Choose a $\beta \in (0,1)$
  - Start with $t = 1$
  - While $f(\mathbf{x} + t\Delta\mathbf{x}) > f(\mathbf{x}) + c_1 t \nabla^T f(\mathbf{x}) \Delta\mathbf{x}$, do
    - ★ Update $t \leftarrow \beta t$
- On convergence, $f(\mathbf{x} + t\Delta\mathbf{x}) \leq f(\mathbf{x}) + c_1 t \nabla^T f(\mathbf{x}) \Delta\mathbf{x}$
- For gradient descent, this means $f(\mathbf{x} + t\Delta\mathbf{x}) \leq f(\mathbf{x}) - c_1 t \|\nabla f(\mathbf{x})\|^2$
- For a function f with Lipschitz continuous $\nabla f(\mathbf{x})$ we have that $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{\hat{t}}{2} \left\|\nabla f(\mathbf{x}^k)\right\|^2$ is satisfied if $\hat{t} = min\left\{1, \beta \frac{2(1-c_1)}{L}\right\}$
- Reason: With backtracking step seach, if $1 - \frac{Lt^k}{2} \geq c_1$, the Armijo rule will be satisfied. That is, $0 < t^k \leq \frac{2(1-c_1)}{L}$

Aside: Backtracking ray search and Lipschitz Continuity

- Recap the Backtracking ray search algorithm
  - Choose a $\beta \in (0, 1)$
  - Start with $t = 1$
  - While $f(\mathbf{x} + t\Delta\mathbf{x}) > f(\mathbf{x}) + c_1 t \nabla^T f(\mathbf{x})\Delta\mathbf{x}$, do
    - ★ Update $t \leftarrow \beta t$
- On convergence, $f(\mathbf{x} + t\Delta\mathbf{x}) \leq f(\mathbf{x}) + c_1 t \nabla^T f(\mathbf{x})\Delta\mathbf{x}$
- For gradient descent, this means $f(\mathbf{x} + t\Delta\mathbf{x}) \leq f(\mathbf{x}) - c_1 t \|\nabla f(\mathbf{x})\|^2$
- For a function f with Lipschitz continuous $\nabla f(\mathbf{x})$ we have that $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{\widehat{t}}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$ is satisfied if $\widehat{t} = min\left\{1, \beta\frac{2(1-c_1)}{L}\right\}$
- Reason: With backtracking step seach, if $1 - \frac{Lt^k}{2} \geq c_1$, the Armijo rule will be satisfied. That is, $0 < t^k \leq \frac{2(1-c_1)}{L} \implies 1 - \frac{Lt^k}{2} \geq c_1$. If not, there must exist an integer j for which $\beta\frac{2(1-c_1)}{L} \leq \beta^j \leq \frac{2(1-c_1)}{L}$, we take $\widehat{t} = min\left\{1, \beta\frac{2(1-c_1)}{L}\right\}$

Rates of Convergence

Convergence Order of convergence (generally Q convergence)



R (root) convergence and Q (quotient) convergence..

R-convergence

- Let us consider the convergence result we got by assuming Lipschitz continuity with backtracking and exact line searches:

$$f(x^k) - f(x^*) \leq \frac{\left\| x^{(0)} - x^* \right\|^2}{2tk}$$

- We will characterize this using **R-convergence**
- 'R' here stands for 'root', as we are looking at convergence rooted at x^*

Q-convergence

- We say that the sequence $s^1, \ldots, s^k$ is **R-linearly** convergent if $\left\| s^k - s^* \right\| \leq v^k, \forall k$, and $\left\{ v^k \right\}$ converges **Q-linearly** to zero
- $v^1, \ldots, v^k$ is Q-linearly convergent if

$$\frac{\left\| v^{k+1} - v^* \right\|}{\left\| v^k - v^* \right\|} \leq r$$

for some $k \geq \theta$, and $r \in (0, 1)$
  - 'Q' here stands for 'quotient' of the norms as shown above

R-convergence assuming Lipschitz continuity

- Consider $v^k = \frac{\left\|x^{(0)} - x^*\right\|^2}{2tk} = \frac{\alpha}{k}$, where α is a constant
- Here, we have $\frac{\left\|v^{k+1} - v^*\right\|}{\left\|v^k - v^*\right\|} \leq \frac{K}{K+1}$, where K is the final number of iterations
  - $\frac{K}{K+1} < 1$, but we don't have $\frac{K}{K+1} < r$
- Thus, $v^k = \frac{\alpha}{k}$ is not Q-linearly convergent as there exist no $v < 1$ s.t. $\frac{\alpha/(k+1)}{\alpha/k} = \frac{k}{k+1} \leq v, \forall k \geq \theta$
- Strictly speaking, for Lipschitz continuity alone, gradient descent is not guaranteed to give R-linear convergence
- In practice, Lipschitz continuity gives "almost" R-linear convergence – not too bad!
- We say that gradient descent with Lipschtiz continuity has convergence rate $O(1/k)$, that is,

R-convergence assuming Lipschitz continuity

- Consider $v^k = \frac{\left\|x^{(0)} - x^*\right\|^2}{2tk} = \frac{\alpha}{k}$, where α is a constant
- Here, we have $\frac{\left\|v^{k+1} - v^*\right\|}{\left\|v^k - v^*\right\|} \leq \frac{K}{K+1}$, where K is the final number of iterations
  - $\frac{K}{K+1} < 1$, but we don't have $\frac{K}{K+1} < r$
- Thus, $v^k = \frac{\alpha}{k}$ is not Q-linearly convergent as there exist no $v < 1$ s.t. $\frac{\alpha/(k+1)}{\alpha/k} = \frac{k}{k+1} \leq v, \forall k \geq \theta$
- Strictly speaking, for Lipschitz continuity alone, gradient descent is not guaranteed to give R-linear convergence
- In practice, Lipschitz continuity gives "almost" R-linear convergence – not too bad!
- We say that gradient descent with Lipschtiz continuity has convergence rate $O(1/k)$, that is, to obtain $f(\mathbf{x}^k) - f(\mathbf{x}^*) \leq \epsilon$, we need $O(\frac{1}{\epsilon})$ iterations.

- Taking hint from this analysis, if Q-linear,

$$\frac{\left\|s^{k+1} - s^*\right\|}{\left\|s^k - s^*\right\|} \leq r \in (0,1)$$

then,

$$\left\|s^{k+1} - s^*\right\| \leq r\left\|s^k - s^*\right\|$$

$$\leq r^2\left\|s^{k-1} - s^*\right\|$$

$$\vdots$$

$\leq r^k\left\|s^{(0)} - s^*\right\|$, which is v^k for R-linear

We want to show convergence that is as specific as possible

- Thus, Q-linear convergence $\Longrightarrow$ R-linear convergence
  - Q-linear is a special case of R-linear
  - R-linear gives a more general way of characterizing linear convergence
- Q-linear is an 'order of convergence'

  r is the 'rate of convergence'

- Q-superlinear convergence:

$$\lim_{k\to\infty} \frac{\left\| s^{k+1} - s^* \right\|}{\left\| s^k - s^* \right\|^2} = 0$$

- Q-sublinear convergence:

$$\lim_{k\to\infty} \frac{\left\| s^{k+1} - s^* \right\|}{\left\| s^k - s^* \right\|} = 1$$

  - *e.g.* For Lipschitz continuity, v^k in gradient descent is Q-sublinear: $\lim_{k\to\infty} \frac{k}{k+1} = 1$

- Q-convergence of order p:

$$\forall k \geq \theta, \frac{\left\| s^{k+1} - s^* \right\|}{\left\| s^k - s^* \right\|^p} \leq M$$

  - *e.g.* $p = 2$ for Q-quadratic, $p = 3$ for Q-cubic, *etc.*
  - M is called the asymptotic error constant

Illustrating Order Convergence

Claim: Every element of s2 appears in s1

Expect s2 to converge faster

- Consider the two sequences $\mathbf{s}_1$ and $\mathbf{s}_2$.

  $\mathbf{s}_1 = \left[\frac{11}{2}, \frac{21}{4}, \frac{41}{8}, \ldots, 5 + \frac{1}{2^n}, \ldots\right]$

  $\mathbf{s}_2 = \left[\frac{11}{2}, \frac{41}{8}, \frac{641}{128}, \ldots, 5 + \frac{1}{2^{2^n-1}}, \ldots\right]$

  Both sequences converge to 5. However, it seems that the second converges faster to 5 than the first one.

- For $\mathbf{s}_1$, $s_1^* = 5$ and Q-convergence is of order $p = 1$ because:

$$\frac{\left\|s_1^{k+1} - s_1^*\right\|}{\left\|s_1^k - s_1^*\right\|^1} = \frac{\left\|\frac{1}{2^{k+1}}\right\|}{\left\|\frac{1}{2^k}\right\|} = \frac{1}{2} < 0.6(= M)$$

- For $\mathbf{s}_2$, $s_2^* = 5$ and Q-convergence is of order $p = 2$ because:

$$\frac{\left\|s_2^{k+1} - s_2^*\right\|}{\left\|s_2^k - s_2^*\right\|^2} = \frac{\left\|\frac{1}{2^{2^{k+1}-1}}\right\|}{\left\|\frac{1}{2^{2^k-1}}\right\|^2} = \frac{1}{2} < 0.6(= M)$$

- **Claim:** Q-convergences of the order p are special cases of Q-superlinear convergence
- $\forall k \geq \theta$, $\frac{\left\| s^{k+1} - s^* \right\|}{\left\| s^k - s^* \right\|^p} \leq M$

$$\implies \lim_{k \to \infty} \frac{\left\| s^{k+1} - s^* \right\|}{\left\| s^k - s^* \right\|} \leq \lim_{k \to \infty} M \left\| s^k - s^* \right\|^{p-1} = 0$$

- Therefore, irrespective of the value of M (as long as $M \geq 0$), order $p > 1$ implies Q-superlinear convergence

Homework?

Question: Could we analyze Gradient descent more **specifically**?

- Assume backtracking line search
- Continue assuming Lipschitz continuity
  - Curvature is upper bounded: $\nabla^2 f(x) \preceq LI$
- Assume **strong convexity**
  - Curvature is lower bounded: $\nabla^2 f(x) \succeq mI$
  - For instance, we might not want to use gradient descent for a quadratic function (curvature is not accounted for)

There exits (Fenchel) duality between strong convexity and Lipschitz continuous gradient. That is, with a good understanding of one, we can easily understand the other one. See `http://xingyuzhou.org/blog/notes/Lipschitz-gradient` for a quick summary!

(Better) Convergence Using Strong Convexity

 Homework: Understand proofs

Theorem

A twice differential function $f : \mathcal{D} \rightarrow \Re$ *for a nonempty open convex set* $\mathcal{D}$

1. *is convex if and only if its domain is convex and its Hessian matrix is positive semidefinite at each point in* $\mathcal{D}$. *That is* $\nabla^2 f(\mathbf{x}) \succeq 0 \quad \forall\ \mathbf{x} \in \mathcal{D}$
2. *is strictly convex if its domain is convex and its Hessian matrix is positive definite at each point in* $\mathcal{D}$. *That is* $\nabla^2 f(\mathbf{x}) \succ 0 \quad \forall\ \mathbf{x} \in \mathcal{D}$
3. *is uniformly convex if and only if its domain is convex and its Hessian matrix is uniformly positive definite at each point in* $\mathcal{D}$. *That is, for any* $\mathbf{v} \in \Re^n$ *and any* $\mathbf{x} \in \mathcal{D}$, *there exists a* $c > 0$ *such that* $\mathbf{v}^T \nabla^2 f(\mathbf{x}) \mathbf{v} \geq c||\mathbf{v}||^2$

Proof of Second Order Conditions for Convexity

In other words

$$\nabla^2 f(\mathbf{x}) \succeq cI_{n\times n}$$

where $I_{n\times n}$ is the $n \times n$ identity matrix and $\succeq$ corresponds to the positive semidefinite inequality. That is, the function f is strongly convex *iff* $\nabla^2 f(\mathbf{x}) - cI_{n\times n}$ is positive semidefinite, for all $\mathbf{x} \in \mathcal{D}$ and for some constant $c > 0$, which corresponds to the positive minimum curvature of f.

PROOF: We will prove only the first statement; the other two statements are proved in a similar manner.

Necessity: Suppose f is a convex function, and consider a point $\mathbf{x} \in \mathcal{D}$. We will prove that for any $\mathbf{h} \in \Re^n$, $\mathbf{h}^T \nabla^2 f(\mathbf{x})\mathbf{h} \geq 0$. Since f is convex, we have

$$f(\mathbf{x} + t\mathbf{h}) \geq f(\mathbf{x}) + t\nabla^T f(\mathbf{x})\mathbf{h} \tag{46}$$

Consider the function $\phi(t) = f(\mathbf{x} + t\mathbf{h})$ defined on the domain $\mathcal{D}_\phi = [0, 1]$.

Proof of Second Order Conditions for Convexity (contd.)

Using the chain rule,

$$\phi'(t) = \sum_{i=1}^{n} f_{x_i}(\mathbf{x} + t\mathbf{h})\frac{dx_i}{dt} = \mathbf{h}^T.\nabla f(\mathbf{x} + t\mathbf{h})$$

Since f has partial and mixed partial derivatives, ϕ' is a differentiable function of t on $\mathcal{D}_\phi$ and

$$\phi''(t) = \mathbf{h}^T\nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

Since ϕ and ϕ' are continous on $\mathcal{D}_\phi$ and ϕ' is differentiable on $int(\mathcal{D}_\phi)$, we can make use of the Taylor's theorem with $n = 3$ to obtain:

$$\phi(t) = \phi(0) + t.\phi'(0) + t^2.\frac{1}{2}\phi''(0) + O(t^3)$$

Writing this equation in terms of f gives

Proof of Second Order Conditions for Convexity (contd.)

Using the chain rule,

$$\phi'(t) = \sum_{i=1}^{n} f_{x_i}(\mathbf{x} + t\mathbf{h})\frac{dx_i}{dt} = \mathbf{h}^T.\nabla f(\mathbf{x} + t\mathbf{h})$$

Since f has partial and mixed partial derivatives, ϕ' is a differentiable function of t on $\mathcal{D}_\phi$ and

$$\phi''(t) = \mathbf{h}^T\nabla^2 f(\mathbf{x} + t\mathbf{h})\mathbf{h}$$

Since ϕ and ϕ' are continous on $\mathcal{D}_\phi$ and ϕ' is differentiable on $int(\mathcal{D}_\phi)$, we can make use of the Taylor's theorem with $n = 3$ to obtain:

$$\phi(t) = \phi(0) + t.\phi'(0) + t^2.\frac{1}{2}\phi''(0) + O(t^3)$$

Writing this equation in terms of f gives

$$f(\mathbf{x} + t\mathbf{h}) = f(\mathbf{x}) + t\mathbf{h}^T\nabla f(\mathbf{x}) + t^2\frac{1}{2}h^T\nabla^2 f(\mathbf{x})\mathbf{h} + O(t^3)$$

Proof of Second Order Conditions for Convexity (contd.)

In conjunction with (46), the above equation implies that

$$\frac{t^2}{2} h^T \nabla^2 f(\mathbf{x})\mathbf{h} + O(t^3) \geq 0$$

Dividing by t^2 and taking limits as $t \to 0$, we get

$$h^T \nabla^2 f(\mathbf{x})\mathbf{h} \geq 0$$

Proof of Second Order Conditions for Convexity (contd.)

Sufficiency: Suppose that the Hessian matrix is positive semidefinite at each point $\mathbf{x} \in \mathcal{D}$. Consider the same function $\phi(t)$ defined above with $\mathbf{h} = \mathbf{y} - \mathbf{x}$ for $\mathbf{y}, \mathbf{x} \in \mathcal{D}$. Applying Taylor's theorem with $n = 2$ and $a = 0$, we obtain,

$$\phi(1) = \phi(0) + t.\phi'(0) + t^2.\frac{1}{2}\phi''(c)$$

for some $c \in (0, 1)$. Writing this equation in terms of f gives

$$f(\mathbf{x}) = f(\mathbf{y}) + (\mathbf{x} - \mathbf{y})^T \nabla f(\mathbf{y}) + \frac{1}{2}(\mathbf{x} - \mathbf{y})^T \nabla^2 f(\mathbf{z})(\mathbf{x} - \mathbf{y})$$

where $\mathbf{z} = \mathbf{y} + c(\mathbf{x} - \mathbf{y})$. Since $\mathcal{D}$ is convex, $\mathbf{z} \in \mathcal{D}$. Thus, $\nabla^2 f(\mathbf{z}) \succeq 0$. It follows that

$$f(\mathbf{x}) \geq f(\mathbf{y}) + (\mathbf{x} - \mathbf{y})^T \nabla f(\mathbf{y})$$

By a previous result, the function f is convex. □