Convergence of Descent Algorithms: Generic and Specific Cases

Back to: Generic Convergence of Descent Algorithm

- Consider the general descent algorithm ($\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$ for each k) with each step: $\mathbf{x}^{k+1} = \mathbf{x}^k + t^k\Delta\mathbf{x}^k$.
  - Suppose f is bounded below in $\Re^n$ and
  - is continuously differentiable in an open set $\mathcal{N}$ containing the level set $\{\mathbf{x} | f(\mathbf{x}) \le f(\mathbf{x}^0)\}$
  - ∇f is Lipschiz continuous.

  Then, $\sum_{k=1}^{\infty} \frac{(\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k)^2}{\|\Delta\mathbf{x}^k\|^2} < \infty$ (that is, it is finite)

  Overall: Sum of squares of normalized directional derivatives is finite

  Proof: normalized directional derivative

- For any descent algorithm: $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$ for each k with each step: $\mathbf{x}^{k+1} = \mathbf{x}^k + t^k\Delta\mathbf{x}^k$.
- From the second Strong Wolfe condition:

Back to: Generic Convergence of Descent Algorithm

- Consider the general descent algorithm ($\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$ for each k) with each step: $\mathbf{x}^{k+1} = \mathbf{x}^k + t^k \Delta\mathbf{x}^k$.
  - Suppose f is bounded below in $\Re^n$ and
  - is continuously differentiable in an open set $\mathcal{N}$ containing the level set $\{\mathbf{x} | f(\mathbf{x}) \le f(\mathbf{x}^0)\}$
  - ∇f is Lipschiz continuous.

  Then, $\displaystyle\sum_{k=1}^{\infty} \frac{(\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k)^2}{\|\Delta\mathbf{x}^k\|^2} < \infty$ (that is, it is finite)

  Proof:

- For any descent algorithm: $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$ for each k with each step: $\mathbf{x}^{k+1} = \mathbf{x}^k + t^k \Delta\mathbf{x}^k$.
- From the second Strong Wolfe condition:

$$\left|\nabla^T f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k)\Delta\mathbf{x}^k\right| \le c_2 \left|\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k\right| \tag{38}$$

Proving Convergence of Descent Algorithm

- Since $c_2 > 0$ and $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$,

Proving Convergence of Descent Algorithm

- Since $c_2 > 0$ and $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$,

$$\nabla^T f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k)\Delta\mathbf{x}^k \geq c_2\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \tag{39}$$

- Subtracting $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$ from both sides of (39)

$$\left[\nabla f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k) - \nabla f(\mathbf{x}^k)\right]^T \Delta\mathbf{x}^k \geq (c_2 - 1)\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \tag{40}$$

- By Cauchy Shwarz inequality and from Lipschitz continuity,

Proving Convergence of Descent Algorithm

- Since $c_2 > 0$ and $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$,

$$\nabla^T f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k)\Delta\mathbf{x}^k \geq c_2\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \tag{39}$$

- Subtracting $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$ from both sides of (39)

$$\left[\nabla f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k) - \nabla f(\mathbf{x}^k)\right]^T \Delta\mathbf{x}^k \geq (c_2 - 1)\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \tag{40}$$

- By Cauchy Shwarz inequality and from Lipschitz continuity,

$$\left[\nabla f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k) - \nabla f(\mathbf{x}^k)\right]^T \Delta\mathbf{x}^k \leq \|\nabla f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k) - \nabla f(\mathbf{x}^k)\|\|\Delta\mathbf{x}^k\| \leq L\|\Delta\mathbf{x}^k\|^2 t^k \tag{41}$$

Proving Convergence of Descent Algorithm (contd.)

- Combining (40) and (41),

$$t^k \geq \frac{c_2 - 1}{L} \frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|^2} \qquad (42)$$

- Substituting (42) into the first Wolfe condition (while recalling that $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$),

Proving Convergence of Descent Algorithm (contd.)

- Combining (40) and (41),

$$t^k \geq \frac{c_2 - 1}{L} \frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|^2} \tag{42}$$

- Substituting (42) into the first Wolfe condition (while recalling that $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$), $f(\mathbf{x}^k + t\Delta\mathbf{x}^k) < f(\mathbf{x}^k) + c_1 t \nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$

$$f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k) - c_1 \frac{1 - c_2}{L} \frac{\left(\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k\right)^2}{\|\Delta\mathbf{x}^k\|^2} \tag{43}$$

- Substituting $c = c_1 \frac{1-c_2}{L}$ and applying (43) successively,

Proving Convergence of Descent Algorithm (contd.)

- Combining (40) and (41),

$$t^k \geq \frac{c_2 - 1}{L} \frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|^2} \tag{42}$$

- Substituting (42) into the first Wolfe condition (while recalling that $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$), $f(\mathbf{x}^k + t\Delta\mathbf{x}^k) < f(\mathbf{x}^k) + c_1 t\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$

$$f(\mathbf{x}^{k+1}) < f(\mathbf{x}^k) - c_1\frac{1-c_2}{L}\frac{\left(\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k\right)^2}{\|\Delta\mathbf{x}^k\|^2} \tag{43}$$

- Substituting $c = c_1\frac{1-c_2}{L}$ and applying (43) successively,

c>0

$$f(\mathbf{x}^{k+1}) < f(\mathbf{x}^0) - c\sum_{i=0}^{k}\frac{\left(\nabla^T f(\mathbf{x}^i)\Delta\mathbf{x}^i\right)^2}{\|\Delta\mathbf{x}^i\|^2} \tag{44}$$

Proving Convergence of Descent Algorithm (contd.)

- Taking limits of (44) as $k \to \infty$,

$$\lim_{k\to\infty} c \sum_{i=0}^{k} \frac{\left(\nabla^T f(\mathbf{x}^i)\Delta\mathbf{x}^i\right)^2}{\|\Delta\mathbf{x}^i\|^2} < \lim_{k\to\infty} f(\mathbf{x}^0) - f(\mathbf{x}^{k+1}) \leq \infty \tag{45}$$

where the last inequality is because the descent algorithm proceeds only if $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k)$, and we have assumed that f is bounded below in $\Re^n$. This proves finiteness of the summation

- Thus, $\lim_{k\to\infty} \frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|} = 0$

[5] Making use of the Cauchy Schwarz inequality

Proving Convergence of Descent Algorithm (contd.)

- Taking limits of (44) as $k \to \infty$,

$$\lim_{k\to\infty} c\sum_{i=0}^{k} \frac{\left(\nabla^T f(\mathbf{x}^i)\Delta\mathbf{x}^i\right)^2}{\|\Delta\mathbf{x}^i\|^2} < \lim_{k\to\infty} f(\mathbf{x}^0) - f(\mathbf{x}^{k+1}) \leq \infty \tag{45}$$

where the last inequality is because the descent algorithm proceeds only if $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k)$, and we have assumed that f is bounded below in $\Re^n$. This proves finiteness of the summation

- Thus, $\lim_{k\to\infty} \frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|} = 0$.
- If we additionally assume that the descent direction is **never** orthogonal to the gradient, *i.e.*, $-\frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|\|\nabla f(\mathbf{x}^k)\|} \geq \Gamma$ for some $\Gamma > 0$, then, we can show[5] that

[5] Making use of the Cauchy Schwarz inequality

Proving Convergence of Descent Algorithm (contd.)

- Taking limits of (44) as $k \to \infty$,

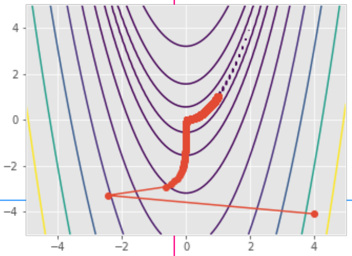
$$\lim_{k\to\infty} c \sum_{i=0}^{k} \frac{\left(\nabla^T f(\mathbf{x}^i)\Delta\mathbf{x}^i\right)^2}{\|\Delta\mathbf{x}^i\|^2} < \lim_{k\to\infty} f(\mathbf{x}^0) - f(\mathbf{x}^{k+1}) \leq \infty \tag{45}$$

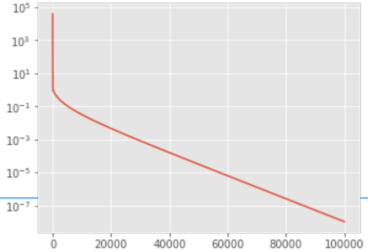
where the last inequality is because the descent algorithm proceeds only if $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k)$, and we have assumed that f is bounded below in $\Re^n$. This proves finiteness of the summation

- Thus, $\lim\limits_{k\to\infty} \dfrac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|} = 0.$
- If we additionally assume that the descent direction is **never** orthogonal to the gradient, *i.e.*, $-\frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|\|\nabla f(\mathbf{x}^k)\|} \geq \Gamma$ for some $\Gamma > 0$, then, we can show[5] that $\lim\limits_{k\to 0} \|\nabla f(\mathbf{x}^k)\| = 0$
- This shows convergence for a generic descent algorithm. What we are more interested in however, is the **rate of convergence of specific** descent algorithms.

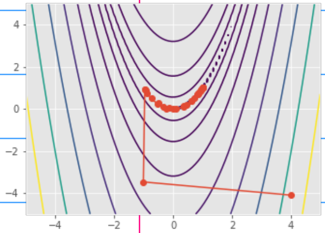
nothing about for what k?

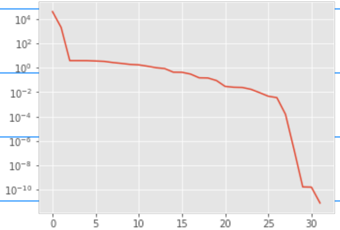
[5] Making use of the Cauchy Schwarz inequality





We wil first look at the rate of convergence of GRADIENT DESCENT for convex functions under Strong Wolfe conditions, Lipschitz continuity on the gradient





We desire the second rate of convergence
But to discuss rate of convergence (as agains an abstract notion of convergence), we will nee to assume
a) convexity and b) specifi form of descent algorithm

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 (36) 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 (36) 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 desire to have the following (46). It holds if....

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{\widehat{t}}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 \tag{46}$$

Convergence of the Gradient Descent Algorithm

- We recap the (necessary) inequality (36) 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 desire to have the following (46). It holds if....

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{\widehat{t}}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 \tag{46}$$

for general descent algos, exact and backtracking sear
for t were motivated

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

For gradient descent with Lipschitz continuity on gradient, here is another way of choosing t

Convergence of the Gradient Descent Algorithm

- We recap the (necessary) inequality (36) 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 desire to have the following (46). It holds if....

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{\widehat{t}}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 \tag{46}$$

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

Convergence of the Gradient Descent Algorithm

- We recap the (necessary) inequality (36) 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 desire to have the following (46). It holds if....

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{\widehat{t}}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 \tag{46}$$

  - 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, (46) 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,
  $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - \frac{t}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$
  $\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^*) + \nabla^\top f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) - \frac{t}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2$
  $\implies f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^*) + \frac{1}{2t}\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2 + \nabla^T f(\mathbf{x}^k)(\mathbf{x}^k - \mathbf{x}^*) - \frac{t}{2}\left\|\nabla f(\mathbf{x}^k)\right\|^2 - \frac{1}{2t}\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2$

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

$$\implies f(\mathbf{x}^{k+1}) - f(\mathbf{x}^*) \leq \frac{1}{2t}\left(\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2 - \left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\|^2\right) \tag{47}$$