1.

Let's find the gradient of the function first.

$$\frac{1}{2}||Ax-b||^2 = \frac{1}{2}\langle Ax-b, Ax-b\rangle$$

$$\frac{1}{2}\left[D_{x_0}\langle Ax-b, Ax-b\rangle\right](h) =$$

$$= \frac{1}{2}\left(\langle\left[D_{x_0}(Ax-b)\right](h), Ax_0-b\rangle + \langle Ax_0-b, \left[D_{x_0}(Ax-b)\right](h)\rangle\right) =$$

$$= \frac{1}{2} * 2\langle Ax_0-b, \left[D_{x_0}(Ax-b)\right](h)\rangle = \frac{1}{2} * 2\langle Ax_0-b, Ah\rangle$$
$$= \langle A^T(Ax_0-b), h\rangle$$

We get: $\nabla_{x_0} f = A^T(Ax_0-b)$

$$\left[D_{x_0}(Ax-b)\right](h)$$

Minimum is obtained when gradient is zero.

$$x_k = A^T b(A^T A)^{-1}$$

In order to be sure that the point is, indeed, minimum, we have to find the second derivative.

$$\left[D_{x_0}\frac{1}{2}||Ax-b||^2\right](h_1) = \langle A^T(Ax_0-b), h_1\rangle$$

Differentiate again.

$$\left[D_{x_0}\langle A^T(Ax_0-b), h_1\rangle\right](h_2) = \langle\left[D_{x_0}(A^T Ax - A^T b)\right](h_2), h_1\rangle = \langle A^T A h_2, h_1\rangle =$$
$$= h_2^T A^T A h_1$$

We found the squared form of second derivative, it does not depend on the point (as initial function is of the second power, therefore the second derivative must be constant.)

In order to show that x_k is the minimum point we need to show that this squared form is positive definite.

$$h^T A^T A h = Ah(Ah)^T = ||Ah||^2 \geq 0$$

It will be zero only when Ah=0.

The only solution when matrix A is zero and b=0, then f(x)=0.

2.

$$f(x) = \frac{1}{2} ||Ax - b||^2$$

Domain: $x \in R^N$.

Want to show all the values that f can obtain.

$$f(x) \geq 0$$

Previously we found that minimum of the function is obtained at $x_k = A^T b (A^T A)^{-1}$,then after substituting:

$$f(x_k) = \frac{1}{2} ||A(A^T b(A^T A)^{-1}) - b||^2 = 0$$

Now let's find maximum of the function:

$$f(x) < f(x_0)$$

If point x_0 is the maximum, then derivative is either zero or does not exist. If $f''(x_0) < 0$, x_0 – is the point of local maximum.

$$f''(x_0) = A^T A = const$$

We get that there is no local maximum, but the global maximum is obtained at x_0, which lies at the end, however, x is defined as all real numbers, therefore, the function does not have local and global maximum.

So, $f(x) \geq 0$, where $x \in R^N$.

3.

Function $f(x)$ satisfies Lipschitz criteria on the interval $[a,b]$ if there exists such $L>0$ that for all x' and $x'' \in [a,b]$.

$$\frac{|f(x')-f(x'')|}{|x'-x''|} \leq L$$

If the inequality is satisfied with constant L, then it is also satisfied for all $L' > L$. For function that satisfies Lipschitz criteria exists infinitely many constants L. Using the algorithms of optimizations that include L as a parameter, best results are obtained when the minimal L is picked.

From $\frac{|f(x\prime)-f(x\prime\prime)|}{|x\prime-x\prime\prime|} \leq L$ follows continuity $f(x)$ on $[a,b]$. According to Weierstrass theorem, $f(x)$ that satisfies on that interval Lipschitz's criteria, has at least 1 point of minimum.

$\frac{|f(x\prime)-f(x\prime\prime)|}{|x\prime-x\prime\prime|} \leq L$ this inequality means that the absolute value of slope coeff. of any horde does not exceed L. When $x'-x'' \to 0$ converges, we see that if at some point there exists tangent line to $f(x)$, then its absolute value of slope coeff. Also does not exceed L. If $f(x)$ has continuous derivative on $[a,b]$, then it satisfies Lipschitz's criteria on that interval with constant $L = \max|f(x')|$=$\max \sigma(A^TA)$, where σ(·) – spectrum (set of eigenvalues).

4.

The stopping criteria in gradient descent method is based on gradient's norm:

$$||\nabla f(x_\mathrm{k})||_2^2 \leq \varepsilon$$

It is squared because for smooth function the difference $f(x_\mathrm{k}) - \min(f)$ implies $||\nabla f(x_\mathrm{k})||_2^2$, not$||\nabla f(x_\mathrm{k})||_2^1$

In order for criteria not to change when f goes to tf, where t>0, it is worth using the following variant of criteria:

$$||\nabla f(x_\mathrm{k})||_2^2 \leq \varepsilon_1 ||\nabla f(x_0)||_2^2$$

where $\varepsilon_1 \in (0,1]$ – chosen relative accuracy.

In such way, stopping criteria guarantees that method will reduce starting error of $\|\nabla f(x_0)\|_2 \quad {\varepsilon_1}^{-1}$ times more.

As $\min(f) = 0$, then $f(x_\varepsilon) \leq \varepsilon$

N – number of iterations required to find point $x_{\boldsymbol{\varepsilon}}$

L – Lipschitz constant, $L = \max \sigma(A^T A)$, where $\sigma(\cdot)$ – spectrum of a given matrix.

x_{min} – x for $\min(f)$

$$f(x_{k+1}) = f(x_k) - \frac{1}{L} \nabla_{x_{\mathbf{k}}} f$$

$$x_{k+1} = \beta_k\big(f(x_{k+1}) - f(x_k)\big) + \gamma_k(f(x_{k+1}) - x_k)$$

$$\|\nabla f(x_{\mathrm{N}})\|^2 \leq \frac{4L(f(x_0) - f(x_{min}))}{\mathrm{N}^2}$$

From above we get:

$$f(x_{\mathrm{N}}) - \min(f) \leq \frac{L\|x_0 - x_{min}\|^2}{\mathrm{N}^2}$$

$$\|\nabla f(x_{\mathrm{N}})\|^2 \leq \frac{4L}{\mathrm{N}^2} \frac{1}{2\varepsilon} \|\nabla f(x_0)\|^2$$

Thus:

$$\|\nabla f(x_{\mathrm{N}})\| \leq \sqrt{\frac{2L}{\varepsilon \mathrm{N}^2}} \|\nabla f(x_0)\|$$

$$\mathrm{N} = 2\sqrt{\frac{2L}{\varepsilon}} = 2\sqrt{\frac{2\max \sigma(A^T A)}{\varepsilon}}$$

$2\sqrt{\frac{2\max\sigma(A^TA)}{\varepsilon}}$ – this number of iterations is required to find point x_ε such that $f(x_\varepsilon) - \min(f) \le \varepsilon$

5.

$$\|\nabla f(x)\|^2 \ \ge 2[\min\sigma(A^TA)](f(x) - \min(f))$$

where $\sigma(\cdot)$ – spectrum of a given matrix.

1) $\|\nabla f(x)\|^2 \ = \|A^T(Ax-b)\|^2 \ = \ \|A^T\|^2\,\|Ax-b\|^2$

2) As min(f)=0, then

$$2[\min\sigma(A^TA)](f(x) - \min(f)) = 2[\min\sigma(A^TA)]f(x) =$$

$$= 2[\min\sigma(A^TA)]\frac{1}{2}\|Ax-b\|^2 \ = \min\sigma(A^TA)\,\|Ax-b\|^2$$

3) $\|A^T\| = \|A^T\|\frac{\|x\|}{\|x\|} = \frac{\|A^Tx\|}{\|x\|}$

$$\inf_{x\in R^M}\frac{\|A^Tx\|}{\|x\|} \ge \min\sigma(A^TA)$$

As $\inf_{x\in R^N} C$ – the greatest lower bound of C , but $\min \mathrm{C}$ – is the minimum of all numbers bounding C.

Let

$$\inf_{x\in R^M}\frac{\|A^Tx\|}{\|x\|} = \min\sigma(A^TA)$$

4) Then substituting 1,2,3 in the initial inequality, we get:

$$(\min\sigma(A^TA))^{\,2}\,\|Ax-b\|^2 \ \ge \min\sigma(A^TA)\,\|Ax-b\|^2$$

It means that minimum squared is greater or equal to minimum, which indeed is true for all $x \in R^N$.

6.

No, it does not change the response.

N**7.**

The function has continuous 2nd derivative on $[a, b]$. If on this interval 2nd derivative is positive, then function is convex at every point of interval. If on this interval 2nd derivative is negative, then function is concave at every point of interval.

$$f_\mu(x) = f(x) + \frac{\mu}{2}\|x\|^2 = \frac{1}{2}\|Ax - b\|^2 + \frac{\mu}{2}\|x\|^2$$

$$f_\mu'(x) = A^T(Ax - B) + \mu x$$

$$f_\mu''(x) = A^T A + \mu = const$$

Function $f_\mu(x)$ can be called strongly convex if there exists constant $q > 0$ such that for any $x \in R^N$ and $\alpha \in [0,1]$ the following inequality is satisfied:

$$f_\mu[\alpha x + (1 - \alpha)y] \leq \alpha f_\mu(x) + (1 - \alpha)f_\mu(y) - \alpha(1 - \alpha)q\|x - y\|^2$$

$$f_\mu(x) = \frac{1}{2}\|Ax - b\|^2 + \frac{\mu}{2}\|x\|^2$$

Strongly convex follows from equality:

$$f_\mu[\alpha x + (1 - \alpha)y] = \alpha f_\mu(x) + (1 - \alpha)f_\mu(y) - \alpha(1 - \alpha)q\langle x - y, A(x - y)\rangle$$

As $\langle x - y, A(x - y)\rangle \geq \sigma\|x - y\|^2$

where σ – the least eigenvalue of matrix A.

$$\nabla_{x_0} f_\mu = A^T(Ax_0 - b) + \mu x_0$$

In our case the global minimum is obtained at the point where gradient is zero.

At $x_k = A^T b(A^T A + \mu)^{-1}$ the global minimum of the function will be obtained.

8.

$$\left\|\nabla f_\mu(x)\right\|^2 \geq 2\mu(f_\mu(x) - \min(f_\mu))$$

1) $\left\|\nabla f_\mu(x)\right\|^2 = \|A^T(Ax-b)+\mu x\|^2$

$$\left\|\nabla f_\mu(x_\mathrm{N})\right\|^2 = \frac{2\max\sigma(A^TA+\mu)}{\varepsilon\mathrm{N}^2}\left\|\nabla f_\mu(x_0)\right\|^2$$

2) Suppose

$$(f_\mu(x_\mathrm{N})-\min(f_\mu)) = \frac{\max\sigma(A^TA+\mu)\|x_0-x_{min}\|^2}{\mathrm{N}^2}$$

3) Out of properties of strongly convex functions:

$$\|x_0-x_{min}\| \leq \frac{2}{\varepsilon}\left\|\nabla f_\mu(x_0)\right\|$$

Then:

$$\|x_0-x_{min}\|^2 \leq \frac{2}{\varepsilon}\left\|\nabla f_\mu(x_0)\right\|^2$$

4) Summarizing 1, 2 and 3:

$$\left\|\nabla f_\mu(x)\right\|^2 \geq 2\mu\big(f_\mu(x)-\min(f_\mu)\big) \Leftrightarrow$$

$$\Leftrightarrow \frac{2\max\sigma(A^TA+\mu)}{\varepsilon\mathrm{N}^2}\left\|\nabla f_\mu(x_0)\right\|^2 \geq \frac{\max\sigma(A^TA+\mu)\|x_0-x_{min}\|^2}{\mathrm{N}^2}$$

Reduce both sides of inequality by $\frac{\max\sigma(A^TA+\mu)}{\mathrm{N}^2}$

$$\frac{2}{\varepsilon}\left\|\nabla f_\mu(x_0)\right\|^2 \geq \|x_0-x_{min}\|^2$$

It follows that given inequality is true for all $x \in R^N$.

9.

This exercise can be solved similarly to exercise 4, Lipschitz constant is changed:

L – Lipschitz constant, $L = \max\sigma(A^TA+\mu)$, where $\sigma(\cdot)$ – spectrum of a given matrix.

It follows that:

$$\left\|\nabla f_{\mu}(x_{\mathrm{N}})\right\| \leq \sqrt{\frac{2L}{\varepsilon \mathrm{N}^2}}\left\|\nabla f_{\mu}(x_0)\right\|$$

$$\mathrm{N} = 2\sqrt{\frac{2L}{\varepsilon}} = 2\sqrt{\frac{2\max \sigma(A^T A + \mu)}{\varepsilon}}$$

$2\sqrt{\frac{2\max \sigma(A^T A+\mu)}{\varepsilon}}$ – this number of iterations is required to find point x_{ε} , such that $f_{\mu}(x_{\varepsilon}) - \min(f_{\mu}) \leq \varepsilon$