Review Summary: Pass😁

EP3260: Machine Learning Over Networks Homework 1

Stefanos Antaris*[1], Amaru Cuba Gyllensten†[1,2],
Martin Isaksson‡[1,2,3], Sarit Khirirat§[1], and Klas Segeljakt¶[1,2]

[1]*KTH Royal Institute of Technology*
[2]*RISE AI*
[3]*Ericsson Research*

January, 2019

Contents

*antaris@kth.se
†amaru.cuba.gyllensten@ri.se
‡martisak@kth.se
§sarit@kth.se
¶klasseg@kth.se

1 Homework assignment

1.1 a – Strong convexity

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\intercal(\boldsymbol{y}-\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \quad \forall \boldsymbol{x},\boldsymbol{y} \in \mathcal{X}, \mu > 0 \tag{1}$$

Problem 1.1.1. *a)* (1) *is equivalent to a minimum positive curvature*

$$\nabla^2 f(\boldsymbol{x}) \geq \mu \boldsymbol{I}_d,\ \forall \boldsymbol{x} \in \ \mathcal{X} \tag{2}$$

b) (1) *is equivalent to*

$$(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))^\intercal(\boldsymbol{y}-\boldsymbol{x}) \geq \mu\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \tag{3}$$

Proof.

$$\begin{aligned}
(1) &\implies f(\boldsymbol{y}) - f(\boldsymbol{x}) && \geq \nabla f(\boldsymbol{x})^\intercal(\boldsymbol{y}-\boldsymbol{x}) + \frac{\mu}{2}||\boldsymbol{x}-\boldsymbol{y}||_2^2 \\
(1) &\implies f(\boldsymbol{y}) - f(\boldsymbol{x}) && \leq \nabla f(\boldsymbol{y})^\intercal(\boldsymbol{y}-\boldsymbol{x}) - \frac{\mu}{2}||\boldsymbol{x}-\boldsymbol{y}||_2^2 \\
&\implies 0 && \leq (\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))(\boldsymbol{y}-\boldsymbol{x}) - \mu||\boldsymbol{x}-\boldsymbol{y}||_2^2 \\
&\implies (\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))^\intercal(\boldsymbol{y}-\boldsymbol{x}) && \geq \mu||\boldsymbol{x}-\boldsymbol{y}||_2^2 && (4) \\
&\implies (\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))^\intercal\frac{\boldsymbol{y}-\boldsymbol{x}}{||\boldsymbol{y}-\boldsymbol{x}||_2^2} && \geq \mu && (5)
\end{aligned}$$

From (4) we directly have that (1) implies (3). If we let $\boldsymbol{y} \to \boldsymbol{x}$, we see that (5) describes the second directional derivative of f, i.e. $\boldsymbol{v}^T\nabla^2 f(\boldsymbol{x})\boldsymbol{v}$, which is larger than μ for all $\boldsymbol{v}$, from which it follows that (1) implies (2).

Since (5), and therefore (3), follows directly from (2), all that is needed for equivalence is to prove that (2) implies (1). By the second order Taylor theorem we have that there exists a convex combination of $\boldsymbol{x}$ and $\boldsymbol{y}$, $\boldsymbol{z}$ such that

$$\begin{aligned}
&f(\boldsymbol{y}) = f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\intercal(\boldsymbol{y}-\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{y}-\boldsymbol{x})^\intercal\nabla^2 f(\boldsymbol{z})(\boldsymbol{y}-\boldsymbol{x}) \\
(2) &\implies (\boldsymbol{y}-\boldsymbol{x})^\intercal\nabla^2 f(\boldsymbol{z})(\boldsymbol{y}-\boldsymbol{x}) \geq \mu||\boldsymbol{y}-\boldsymbol{z}||_2^2 \qquad\qquad \text{ok} \\
&\implies f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\intercal(\boldsymbol{y}-\boldsymbol{x}) + \frac{\mu}{2}||\boldsymbol{y}-\boldsymbol{x}||_2^2
\end{aligned}$$

■

1.1.1 a — the Polyak-Łojasiewicz inequality

Problem 1.1.2. (1) *implies the Polyak-Łojasiewicz (PL) inequality.*

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leq \frac{1}{2\mu}\left\|\nabla f(\boldsymbol{x})\right\|_2^2,\ \forall \boldsymbol{x} \tag{6}$$

Proof. The right hand term in (1) is convex quadratic w.r.t $\boldsymbol{y}$ and $\boldsymbol{x}$ fixed. We set the gradient with respect to $\boldsymbol{y}$ in (1) to 0 and find that $\tilde{\boldsymbol{y}} = \boldsymbol{x} - \frac{1}{\mu}\nabla f(\boldsymbol{x})$ minimizes the righthand term.

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

ok

This holds for any $\boldsymbol{y} \in S$, therefore also for $\boldsymbol{x}^*$. We arrive at (6) after rearranging the terms. ■

1.1.2 b

Problem 1.1.3. (1) *implies*

$$\|\boldsymbol{y}-\boldsymbol{x}\|_2 \leq \frac{1}{\mu}\left\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\right\|_2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \tag{7}$$

Proof. We can now derive a bound on the distance between two points.

From (3), applying Cauchy-Schwarz's inequality ($\boldsymbol{x}^\mathsf{T}\boldsymbol{y} \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$) yields

$$\begin{aligned} \mu\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 &\leq (\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}))^\mathsf{T}(\boldsymbol{y}-\boldsymbol{x}) \\ \mu\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 &\leq \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|\|\boldsymbol{y}-\boldsymbol{x}\|. \end{aligned}$$

Since $\boldsymbol{y} \neq \boldsymbol{x}$, we can divide both sides by $\|\boldsymbol{y}-\boldsymbol{x}\|$ and conclude that ok

$$\mu\|\boldsymbol{y}-\boldsymbol{x}\| \leq \left\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\right\|.$$

Hence, we complete the proof. ■

1.1.3 c

Problem 1.1.4. (1) *implies*

$$\left(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\right)^\mathsf{T}(\boldsymbol{y}-\boldsymbol{x}) \leq \frac{1}{\mu}\left\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\right\|_2^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \tag{8}$$

Let us introduce one useful lemma which facilitates our proof

Lemma 1.1.1. *Let f be μ-strongly convex. Then,*

$$f(y) \leq f(x) + \nabla f(x)^T(y-x) + \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|^2.$$

Proof. Define $g(y) = f(y) - \nabla f(x)^T y$ for any fixed vector x. Then, we can first prove that g is also $\mu-$strongly convex, since from (3)

ok

$$(\nabla g(y) - \nabla g(x))^T (y - x) = (\nabla f(y) - \nabla f(x))^T (y - x) \geq \mu \|y - x\|^2.$$

(Note that this is the gradient with respect to y). Next, from the definition of the strong convexity of g

$$\begin{aligned} g(y) &\geq g(x) + \nabla g(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2 \\ &\geq \min_y \left[g(x) + \nabla g(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2 \right]. \end{aligned}$$

By the first-order optimality condition with respect to y, we have to set $y = x - \frac{1}{\mu}\nabla g(x)$. Therefore, plugging the result into the main inequality yields

$$g(y) \geq g(x) - \frac{1}{2\mu}\|\nabla g(x)\|^2.$$

This means that

Note:
Applying (6) to g can lead to lemma 1.1.1 directly.

$$g(x) \geq g(y) - \frac{1}{2\mu}\|\nabla g(y)\|^2,$$

or equivalently

$$f(x) - \nabla f(x)^T x \geq f(y) - \nabla f(x)^T y - \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|^2.$$

Therefore,

$$f(x) + \nabla f(x)^T (y - x) + \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|^2 \geq f(y).$$

The proof of lemma 1.1.1 is complete. ■

Proof. Now, we are ready to prove our main result. From lemma 1.1.1 stated above,

$$\begin{aligned} f(\boldsymbol{y}) &\leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\mathsf{T}(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2\mu}\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|^2, \text{ and} \\ f(\boldsymbol{x}) &\leq f(\boldsymbol{y}) + \nabla f(\boldsymbol{y})^\mathsf{T}(\boldsymbol{x} - \boldsymbol{y}) + \frac{1}{2\mu}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2. \end{aligned}$$

Combining two inequalities above yields

$$\left(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\right)^\mathsf{T} (\boldsymbol{x} - \boldsymbol{y}) \leq \frac{1}{\mu}\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|^2.$$

ok

The proof is hence complete.

■

1.1.4 d

Problem 1.1.5. (1) *implies* $f(\boldsymbol{x})+r(\boldsymbol{x})$ *is strongly convex for any convex* f *and strongly convex* r*.*

Proof. Since f is convex and r is strongly convex, the following inequalities hold

$$\begin{aligned} &f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{x}), \text{ and} \\ &r(\boldsymbol{y}) \geq r(\boldsymbol{x}) + \nabla r(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2. \end{aligned}$$

Define $g(\boldsymbol{x}) = f(\boldsymbol{x}) + r(\boldsymbol{x})$. Then, the addition of two inequalities above yield

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2.$$

ok

Therefore, $g(\boldsymbol{x})$ is strongly convex with μ. ■

1.2 b — Smoothness

Lemma 1.2.1. *A function* $f\colon \mathbb{R}^d \to \mathbb{R}$*, is L-smooth iff it is differentiable and its gradient is L-Lipschitz-continuous (usually w.r.t. norm-2).*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d, \left\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\right\|_2 \leq L\|\boldsymbol{y}-\boldsymbol{x}\|_2 \tag{9}$$

1.2.1 a

(9) implies

$$f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \quad \forall \boldsymbol{y}, \boldsymbol{x} \tag{10}$$

Proof. By the continuous differentiability of f and the fundamental theorem of Calculus,

$$\begin{aligned} f(y) &= f(x) + \int_{\tau=0}^{1} \left[\nabla f(x+\tau(y-x))\right]^{\mathsf{T}} (y-x)\, \mathrm{d}\tau \\ &= f(x) + \nabla f(x)^{\mathsf{T}}(y-x) + \int_{\tau=0}^{1} \left[\nabla f(x+\tau(y-x)) - \nabla f(x)\right]^{\mathsf{T}} (y-x)\, \mathrm{d}\tau\,. \end{aligned}$$

Applying Cauchy-Schwarz's inequality (i.e. $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y} \leq \|\boldsymbol{x}\|\|\boldsymbol{y}\|$) yields

OK

Another way: define a function g(x) = L/2XX-f(x)

$$\begin{aligned} f(y) &\leq f(x) + \nabla f(x)^T(y-x) + \int_{\tau=0}^{1} \|\nabla f(x+\tau(y-x)) - \nabla f(x)\|\|y-x\| d\tau \\ &\leq f(x) + \nabla f(x)^T(y-x) + L\int_{\tau=0}^{1} \tau\|y-x\|^2 d\tau \\ &\leq f(x) + \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|^2. \end{aligned}$$

The last second inequality comes from the definition of the smoothness of f. We hence complete the proof. ■

1.2.2 b

(9) implies

$$f(\boldsymbol{x_2}) \geq f(\boldsymbol{x_1}) + \nabla f(\boldsymbol{x_1})^\intercal(\boldsymbol{x_2} - \boldsymbol{x_1}) + \frac{1}{2L}\left\|\nabla f(\boldsymbol{x_2}) - \nabla f(\boldsymbol{x_1})\right\|_2^2 \quad \forall \boldsymbol{x_2}, \boldsymbol{x_1} \tag{11}$$

Proof. Define $g(y) = f(y) - \nabla f(x_0)^T y$, given a fixed vector x_0. Hence, $\nabla g(y) = \nabla f(y) - \nabla f(x_0)$. Then, the optimal point is $y^\star = x_0$ ($\nabla g(y^\star = 0)$), and g is $L-$smooth, i.e.

$$\|\nabla g(x) - \nabla g(y)\| = \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

Since g is $L-$smooth, (10) yields

$$g(x_2) \leq \min_{x_2}\left\{g(x_1) + \nabla g(x_1)^T(x_2 - x_1) + \frac{L}{2}\|x_2 - x_1\|^2\right\}.$$

OK

To minimize the right-hand side of the inequality, we have to set $x_2 = x_1 - (1/L)\nabla g(x_1)$ by the first-order optimality condition with respect to x_1. Therefore, we have:

$$g(x_2) \leq g(x_1) - \frac{1}{2L}\|\nabla g(x_1)\|^2.$$

Plugging $x_2 = x_0$ yields or equivalently

$$f(x_0) - \nabla f(x_0)^T x_0 \leq f(x_1) - \nabla f(x_0)^T x_1 - \frac{1}{2L}\|\nabla f(x_1) - \nabla f(x_0)\|^2.$$

Therefore,

$$f(x_0) + \nabla f(x_0)^T(x_1 - x_0) + \frac{1}{2L}\|\nabla f(x_1) - \nabla f(x_0)\|^2 \leq f(x_1).$$

We thus complete the proof.

■

1.2.3 c

(9) implies

$$\left(\nabla f(\boldsymbol{x_2}) - \nabla f(\boldsymbol{x_1})\right)^\intercal(\boldsymbol{x_2} - \boldsymbol{x_1}) \geq \frac{1}{L}\left\|\nabla f(\boldsymbol{x_2}) - \nabla f(\boldsymbol{x_1})\right\|_2^2, \quad \forall \boldsymbol{x_1}, \boldsymbol{x_2} \tag{12}$$

Proof. From (11),

$$\begin{aligned} f(x) &\geq f(y) + \nabla f(y)^T(x - y) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2, \quad \text{and} \\ f(y) &\geq f(x) + \nabla f(x)^T(y - x) + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2. \end{aligned}$$

OK

Combining two inequalities above yields

$$(\nabla f(y) - \nabla f(x))^T(y - x) \geq \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2.$$

Hence, the proof is complete. ■

1.3 c — Resource allocation

$$\text{minimize } \frac{1}{N}\sum\nolimits_{i\in[N]} f_i(x_i) \text{ s.t } \boldsymbol{Ax} = \boldsymbol{b} \tag{13}$$

for $\boldsymbol{A} \in \mathbb{R}^{p\times N}$ and $\boldsymbol{x} = [x_1, \dots, x_N]^T$

1.3.1 a

Assume strong-convexity and smoothness on f. How would you solve this problem when $N = 1000$?

Solution. Denote $\boldsymbol{A}_i \in \mathbb{R}^{1\times N}$ and $b_i \in \mathbb{R}$ be i^{th} row of $\boldsymbol{A} \in \mathbb{R}^{p\times N}$ and i^{th} element of $b \in \mathbb{R}^p$, respectively. Also, let $\lambda = [\lambda_1, \dots, \lambda_N]^\intercal$ be the dual variable associated with the equality constraint $Ax - b = 0$. The Lagrangian function becomes

$$L(x, \lambda) = \sum_{i\in[N]} \left[\frac{1}{N} f_i(x_i) - \lambda_i(\boldsymbol{A}_i x_i - b_i)\right].$$

ok

Thus, the dual function becomes

$$g(\lambda) = \inf_{x\in\mathbb{R}^N} L(x, \lambda)$$

exists if $\frac{1}{N}\nabla f_i(x_i) - \lambda_i A_i^\intercal = 0$ for $i \in [N]$ (from the first optimality condition with respect to $\boldsymbol{x}$), and $-\infty$ otherwise.

Idea: We can minimize this dual objective using Newton's method.

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \alpha_k \nabla^2 f(\boldsymbol{x}_k)^{-1} \nabla f(\boldsymbol{x}_k) \tag{14}$$

which is feasible because $N = 1000$ and we can reasonably easily compute the Hessian. ■

1.3.2 b

What if $N = 10^9$? We can't compute the Hessian. Idea: Parallelize over i, and perhaps sample

ok

1.3.3 c

Can we use Newton's method for $N = 10^9$? Try efficient method for computing $\nabla^2 f(\boldsymbol{x}_k)$ for $p = 1$ and $b = 1$ (probability simplex constraint). Extend it to $1 \le p \ll N$.

The Hessian can be estimated (quasi-Newton methods) by finite difference of the gradients, which is more efficient). We can also do tricks like re-using the Hessian and only recomputing it every N steps, or replacing it with a diagonal approximation with n second derivatives $\frac{\partial^2 f(x)}{\partial x_i^2}$.

It's a nice idea to re-using Hessian

$\boldsymbol{A} \in \mathbb{R}^{1\times N}$ and $\boldsymbol{Ax} = 1$

1.3.4 d

Now, we add twice differentiable $r(\boldsymbol{x})$ to the objective and solve sections 1.3.1 to 1.3.3.

$$\text{minimize } r(\boldsymbol{x}) + \frac{1}{N} \sum\nolimits_{i\in[N]} f_i(x_i) \text{ s.t } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{15}$$

for $\boldsymbol{A} \in \mathbb{R}^{p\times N}$ and $\boldsymbol{x} = [x_1, \ldots, x_N]^T$

Idea: Not nice since $r(\boldsymbol{x})$ now depends on the entire $\boldsymbol{x}$ and can't be parallelized?

Not sure either

1.4 d — Proof sketch for strongly-convex and L-smooth f

Problem. Let f be μ-strongly convex and $L-$smooth. Then,

OK

$$\left(\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\right)^T (\boldsymbol{y} - \boldsymbol{x}) \geq \frac{\mu L}{\mu + L} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 + \frac{1}{\mu + L} \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2^2.$$

Sol: We begin by introducing the following useful lemmas which facilitates our proof.

Lemma 1.4.1. *Let $g(x) = f(x) - \frac{\mu}{2}\|x\|^2$. Then, g is convex and $(L - \mu)-$smooth.*

Proof. Notice that $\nabla g(x) = \nabla f(x) - \mu x$. We at first prove the convexity of g as follows:

$$\begin{aligned} g(y) &= f(y) - \frac{\mu}{2}\|y\|^2 \\ &\geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2 - \frac{\mu}{2}\|y\|^2 \\ &= g(x) + \nabla g(x)^T (y - x) + \frac{\mu}{2}\|y - x\|^2 - \frac{\mu}{2}\|y\|^2 + \frac{\mu}{2}\|x\|^2 - \mu x^T (x - y), \end{aligned}$$

where the first inequality comes from the strong convexity of f.

Applying the equality $2a^T b = \|a\|^2 + \|b\|^2 - \|a - b\|^2$ with $a = x$ and $b = x - y$ into the main inequality yields

$$g(y) \geq g(x) + \nabla g(x)^T (y - x).$$

Therefore, g is convex. Next, we prove the Lipschitz smoothness of g.

$$\begin{aligned} g(y) &= f(y) - \frac{\mu}{2}\|y\|^2 \\ &\leq f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|y - x\|^2 - \frac{\mu}{2}\|y\|^2 \\ &= g(x) + \nabla g(x)^T (y - x) + \frac{L}{2}\|y - x\|^2 - \frac{\mu}{2}\|y\|^2 + \frac{\mu}{2}\|x\|^2 - \mu x^T (x - y), \end{aligned}$$

where the first inequality comes from the Lipschitz smoothness of f. Applying the equality $2a^T b = \|a\|^2 + \|b\|^2 - \|a - b\|^2$ with $a = x$ and $b = x - y$ into the main inequality yields

$$g(y) \leq g(x) + \nabla g(x)^T (y - x) + \frac{L - \mu}{2}\|y - x\|^2.$$

Therefore, g is $(L - \mu)-$smooth. ■

Now, we are ready to prove the main result. By the coercitivity property of g

$$(\nabla g(y) - \nabla g(x))^T(y - x) \geq \frac{1}{L - \mu}\|\nabla g(y) - \nabla g(x)\|^2.$$

Since $\nabla g(x) = \nabla f(x) - \mu x$, we have:

$$(\nabla f(y) - \nabla f(x))^T(y - x) - \mu\|y - x\|^2 \geq \frac{1}{L - \mu}\|\nabla f(y) - \nabla f(x) - \mu(y - x)\|^2.$$

Plugging

$$\begin{aligned}\|\nabla f(y) - \nabla f(x) - \mu(y - x)\|^2 = {} & \|\nabla f(y) - \nabla f(x)\|^2 \\ & - 2\mu(\nabla f(y) - \nabla f(x))^T(y - x) \\ & + \mu^2\|y - x\|^2\end{aligned}$$

into the main inequality and rearranging the terms yield the result.