Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Department of Computer Science
Bernd Gärtner and Niao He

Optimization for Data Science **Final Exam (19 August 2022)** **FS22**

Candidate

First name: ....

Last name: ....

Student ID (Legi) Nr.: ....

I attest with my signature that I was able to take the exam without any impediments and that I have read and understood the general remarks below.

Signature: ....

General remarks and instructions

1. Check your exam documents for completeness (pages numbered from 1 to 14).
2. Immediately inform an assistant in case you experience any impediments. Complaints after the exam cannot be accepted.
3. You can solve the exercises in any order. They are not ordered by difficulty. Solutions should be written into the provided spaces. If you need scratch paper and/or extra paper for solutions, please ask an assistant.
4. Pencils are not allowed. Pencil-written solutions will not be graded.
5. All electronic devices must be turned off and are not allowed to be on your desk or carried with you to the toilet.
6. Attempts to cheat will be noted and reported to the examination office who will decide on the appropriate legal measures.
7. Provide only one solution to each exercise. Cross out invalid solutions clearly. If multiple solutions are provided, none of them will be graded.
8. **For full points, the explanations for your solutions must be clear, without any gaps, and mathematically rigorous unless explicitly stated otherwise.**
9. You may use (without proof) any statement that has been proved in the lecture and the exercise sessions, appears in previous subtasks of the same assignment, or as a hint in the assignments. If you need something *different* from that, you must write a new proof or at least list all necessary changes.

Good luck!

To be filled by corrector

	achieved points (maximum)
1	(8)
2	(8)
3	(8)
4	(8)
5	(8)
6	(16)
7	(20)
8	(24)
Σ	(100)

Multiple Choices Multiple Answers

Each choice is worth 2 points. You get 2 points for each correct choice you select and 2 points for each incorrect choice you leave blank. For example, suppose a question with four choices A, B, C, and, D. If A and B are the correct choices, then you receive 4 points by selecting A and C (two points for selecting A and two points for not selecting D). Each question has at least one correct choice. No points are awarded if you do not select any choice for a question. You do not need to provide proofs or counterexamples for the choices you select.

Assignment 1 (8 points). *Consider the function* $\phi(x) = \max_{1 \le i \le n} f_i(x)$, *where* $f_i : \mathbb{R}^d \to \mathbb{R}$ *for* $i = 1, \dots, n$. *Which ones of the following statements are* **correct***?*

- ☐ *Function* $\phi(x)$ *is convex if* $f_1, f_2, \dots, f_n$ *are convex.*
- ☐ *Function* $\phi(x)$ *is smooth if* $f_1, f_2, \dots, f_n$ *are smooth.*
- ☐ *Function* $\phi(x)$ *is strongly convex if* $f_1, f_2, \dots, f_n$ *are strongly convex.*
- ☐ *Function* $\phi(x) + \frac{L}{2}\|x\|_2^2$ *is convex if* $f_1, f_2, \dots, f_n$ *are twice continuously differentiable and* $\nabla^2 f_i(x) \succeq -L \cdot I_d$ *for all* $i \in \{1, \dots, n\}$ *and* $x \in \mathbb{R}^d$ *where* I_d *is the identity matrix of size* d.

Solution: A, C, D.

Assignment 2 (8 points). *We say an algorithm is* **affine-invariant** *if the trajectories of the algorithm remain the same when applied to the problem* $\min_{x \in X \subset \mathbb{R}^d} f(x)$ *and to the problem under affine transformation* $\min_{y \in Y} f(Ay)$, *where* $A : \mathbb{R}^d \to \mathbb{R}^d$ *is invertible and* $Y := \{y \in \mathbb{R}^d : Ay \in X\}$. *More specifically, let* $\{x_t\}$ *and* $\{y_t\}$ *be the iterates, respectively; if* $x_0 = Ay_0$, *then one can ensure* $x_t = Ay_t, \forall t \ge 1$.

Which ones of the following methods are affine invariant?

- ☐ *Gradient descent.*
- ☐ *Newton's method.*
- ☐ *Frank-Wolfe method.*
- ☐ *Cubic regularization method.*

Solution: B, C.

Assignment 3 (8 points). *Consider the stochastic optimization problem* $\min_x F(x) := \mathbb{E}_\xi[f(x, \xi)]$. *Let* $\xi_0, \xi_1, \dots, \xi_t$ *be i.i.d. samples and* $x_{t+1} = x_t - \gamma_t \mathbf{g}_t, t \ge 0$, *where* $\mathbf{g}_t$ *is defined recursively below with* $\mathbf{g}_0 = \nabla f(x_0, \xi_0)$. *Which ones of the following (variance-reduced) gradient estimators are* **unconditionally unbiased**, *namely, under total expectation, we have* $\mathbb{E}_{\xi_0, \dots, \xi_t}[\mathbf{g}_t] = \mathbb{E}_{\xi_0, \dots, \xi_t}[\nabla F(x_t)]$ *for any* $t \ge 1$*? Note that this is slightly different from the conditional unbiasness we have seen in the lecture.*

- ☐ $\mathbf{g}_t = \frac{1}{t+1}\nabla f(\mathbf{x}_t, \xi_t) + \frac{t}{t+1}(\mathbf{g}_{t-1} - \nabla f(\mathbf{x}_{t-1}, \xi_t))$
- ☐ $\mathbf{g}_t = \nabla f(\mathbf{x}_t, \xi_t) + \frac{t}{t+1}(\mathbf{g}_{t-1} - \nabla f(\mathbf{x}_{t-1}, \xi_t))$
- ☐ $\mathbf{g}_t = \begin{cases} \nabla f(\mathbf{x}_t, \xi_t) & \textit{with probability } p \\ \mathbf{g}_{t-1} + \nabla f(\mathbf{x}_t, \xi_t) - \nabla f(\mathbf{x}_{t-1}, \xi_t) & \textit{with probability } 1-p \end{cases}$
- ☐ $\mathbf{g}_t = \mathbf{g}_{t-1} + \nabla^2 f(\mathbf{x}_t, \xi_t)(\mathbf{x}_t - \mathbf{x}_{t-1})$

Solution: B, C.

Assignment 4 (8 points). *Consider the minimax problem:* $\min_x \max_y \frac{x^2}{2} + xy - \frac{y^2}{2}$. *Which ones of the following statements are* **correct***?*

- ☐ *Saddle point does not exist.*
- ☐ *Saddle point exists and is unique.*
- ☐ *GDA does not converge under any constant stepsize.*
- ☐ *GDA with small enough constant stepsize converges linearly.*

Solution: B, D.

Assignment 5 (8 points). *Let* $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ *be convex but possibly nonsmooth and admit the global minimizer* $\mathbf{x}^*$. *Which ones of the following statements about* $\mathbf{x}^*$ *are* **correct***?*

- ☐ $\partial f(\mathbf{x}^*) = \{0\}$.
- ☐ $\nabla f_\mu(\mathbf{x}^*) = 0$ *for any* $\mu > 0$, *where* f_μ *is the Moreau envelope of* f.
- ☐ $\mathbf{x}^* = \mathit{prox}_{\mu f}(\mathbf{x}^*)$ *for any* $\mu > 0$, *where* $\mathit{prox}_{\mu f}$ *is the proximal operator of* μf.
- ☐ $\mathbf{x}^*$ *is the global minimizer of* $f^\omega(\mathbf{x}) = \min_{\mathbf{y}}\{f(\mathbf{y}) + V_\omega(\mathbf{y}, \mathbf{x})\}$ *for any Bregman divergence* $V_\omega(\mathbf{y}, \mathbf{x}) := \omega(\mathbf{y}) - \omega(\mathbf{x}) - \langle \nabla\omega(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle$ *with* ω *being convex and continuously differentiable.*

Solution: B, C, D.

Mathematical Proofs

Assignment 6 (16 points). *A function* $f : \mathbb{R}^d \supseteq \mathrm{dom}(f) \to \mathbb{R}_{>0}$ *is called log-convex, if* $\mathrm{dom}(f)$ *is convex and the function* $\log \circ f : \mathrm{dom}(f) \to \mathbb{R}; \mathbf{x} \mapsto \log(f(\mathbf{x}))$ *is convex.*

(a) (4 points) Show that the function $g : \mathbb{R} \to \mathbb{R}_{>0}$ *given by*

$$g(x) = \frac{1 + e^x}{e^x}$$

is log-convex.

(b) (4 points) Show that $f : \mathbb{R}^d \to \mathbb{R}_{>0}$ *is log-convex if and only if*

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq f(\mathbf{x})^{\lambda} f(\mathbf{y})^{1-\lambda} \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \lambda \in [0, 1]$$

(c) (4 points) Let $f, g : \mathbb{R}^d \to \mathbb{R}_{>0}$ *be log-convex. Prove that* fg *and* f^{α} *with* $\alpha \geq 0$ *are log-convex.*

(d) (4 points) Let $f : \mathbb{R}^d \to \mathbb{R}_{>0}$ *be log-convex. Show that* f *is convex.*

Solution:

(a) We start by noticing that the domain of g is $\mathbb{R}$ which is convex and open. By the second-order characterization of convexity and the fact that $\frac{d^2}{dx^2}\log(g(x)) = \frac{e^x}{(1+e^x)^2} > 0$ we conclude that $\log(g)$ is convex and therefore g is log-convex.

(b) Follows by the definition of a convex function applied to $\log(f)$.

(c) In both cases the domain of the functions is $\mathbb{R}^d$ which is convex.(i) $\log((fg)(\mathbf{x})) = \log f(\mathbf{x}) + \log g(\mathbf{x})$ and log-convexity of the product follows by convexity of the sum of two convex functions. (ii) $\log(f^\alpha(\mathbf{x})) = \alpha \log f(\mathbf{x})$ implying log-convexity of f^α.

(d) The domain of f is $\mathbb{R}^d$ which is convex. Consider arbitrary $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\lambda \in [0,1]$,

$$\begin{aligned} e^{\log(f(\lambda \mathbf{x}+(1-\lambda)\mathbf{y})} &\leq e^{\lambda \log(f(\mathbf{x}))+(1-\lambda)\log(f(\mathbf{y}))} \\ &\leq \lambda e^{\log(f(\mathbf{x}))} + (1-\lambda)e^{\log(f(\mathbf{y}))} \\ &= \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}). \end{aligned}$$

Where the first inequality follows from the convexity of $\log(f)$ and the second one from the convexity of the function e^x. We then conclude that f is convex by the definition of convexity.

Assignment 7 (20 points). *A twice continuously differentiable function* $f : \mathbb{R}^d \to \mathbb{R}$ *is called* almost quadratic *if there exist a symmetric matrix* M *and a constant* $\epsilon \in [0, 1)$ *such that* $\|\nabla^2 f(\mathsf{x}) - M\| \leq \epsilon \cdot \underline{\lambda}(M)$ *for all* $\mathsf{x} \in \mathbb{R}^d$*, where* $\underline{\lambda}(M) > 0$ *denotes the smallest of the* absolute *eigenvalues of* M.

Let x, x' *be two consecutive iterates of Newton's method, run on an almost quadratic function* f *with parameter* ϵ.

Prove that

$$\|\mathsf{x}' - \mathsf{x}^*\| \leq \frac{2\epsilon}{1-\epsilon}\|\mathsf{x} - \mathsf{x}^*\|,$$

where $\mathsf{x}^\star$ *is a critical point of* f. *(This implies that Newton's method globally converges if* $\varepsilon < 1/3$*, and that in this case, there is at most one critical point.)*

Hint: *we can follow the analysis of Newton's method from the lecture (proof of Theorem 10.4) and eventually reach the inequality*

$$\|\mathsf{x}' - \mathsf{x}^*\| \leq \|\mathsf{x} - \mathsf{x}^*\| \cdot \|H(\mathsf{x})^{-1}\| \int_0^1 \|H(\mathsf{x} + t(\mathsf{x}^* - \mathsf{x})) - H(\mathsf{x})\| dt.$$

Here, $H(\mathsf{x})$ *is a shortcut for* $\nabla^2 f(\mathsf{x})$.

You may also use that for any symmetric square matrix A *and any unit vector* $\mathbf{v}$,

$$\underline{\lambda}(A) \leq \|A\mathbf{v}\| \leq \overline{\lambda}(A),$$

where $\overline{\lambda}(A)$ *is the largest of the absolute eigenvalues of* A *(both bounds are tight).*

Solution: We first bound $||H(\mathbf{x})^{-1}||$. Eigenvalues of $H(\mathbf{x})^{-1}$ are the inverses of the eigenvalues of $H(\mathbf{x})$, we have that

$$||H(\mathbf{x})^{-1}|| = \frac{1}{\underline{\lambda}(H(\mathbf{x}))},$$

using the hint and definition of the spectral norm.

Let $\mathbf{v}$ be a unit eigenvector of $H(\mathbf{x})$ for eigenvalue $\underline{\lambda}(H(\mathbf{x}))$. Using triangle inequality and properties of the spectral norm, we have

$$\begin{aligned}
\underline{\lambda}(H(\mathbf{x})) &= ||H(\mathbf{x})\mathbf{v}|| \\
&\geq ||M\mathbf{v}|| - ||(H(\mathbf{x}) - M)\mathbf{v}|| \\
&\geq \underline{\lambda}(M) - ||(H(\mathbf{x}) - M)||||\mathbf{v}|| \\
&\geq (1-\epsilon)\underline{\lambda}(M).
\end{aligned}$$

Hence,

$$||H(\mathbf{x})^{-1}|| \leq \frac{1}{(1-\epsilon)\underline{\lambda}(M)}.$$

Next we bound $||H(\mathbf{x} + t(\mathbf{x}^* - \mathbf{x})) - H(\mathbf{x})||$. Let $\mathbf{y} = \mathbf{x} + t(\mathbf{x}^* - \mathbf{x})$.

By triangle inequality for the spectral norm,

$$||H(\mathbf{x}) - H(\mathbf{y})|| \leq ||H(\mathbf{x}) - M|| + ||H(\mathbf{y}) - M|| \leq 2\epsilon\underline{\lambda}(M).$$

Plugging both bounds into the analysis, the conclusion follows.

Assignment 8 (24 points). *Consider the stochastic optimization problem:*

$$\min_{\mathsf{x}} F(\mathsf{x}) = \mathbb{E}_{\xi\sim P}[f_\xi(\mathsf{x})] := \sum_{i=1}^{n} p_i f_i(\mathsf{x}) \tag{1}$$

where $P(\xi = i) = p_i \geq 0, \sum_{i=1}^n p_i = 1$. *Assume that* $f_i(\mathsf{x})$ *is* L_i*-smooth and convex for any* $i = 1, \dots, n$ *and* $F(\mathsf{x})$ *is* μ*-strongly convex. In addition, assume that there exists* x^* *such that* $\nabla f_i(\mathsf{x}^*) = 0, \forall i = 1, \dots, n$ *(interpolation regime).*

(a) *(5 points) Prove that for any* i, x, *it holds that*

$$\langle \mathsf{x} - \mathsf{x}^*, \nabla f_i(\mathsf{x}) \rangle \geq \frac{1}{L_i} ||\nabla f_i(\mathsf{x})||^2,$$

and

$$\langle \nabla F(\mathsf{x}), \mathsf{x} - \mathsf{x}^* \rangle \geq \mu ||\mathsf{x} - \mathsf{x}^*||^2.$$

You can use the following fact that for any L*-smooth and convex function* f, *it holds that*

$$f(\mathbf{y}) \geq f(\mathsf{x}) + \langle \nabla f(\mathsf{x}), \mathbf{y} - \mathsf{x} \rangle + \frac{1}{2L} ||\nabla f(\mathsf{x}) - \nabla f(\mathbf{y})||_2^2, \forall \mathsf{x}, \mathbf{y}.$$

(b) *(6 points) Under the above assumptions, prove that SGD with constant stepsize : for* $t \geq 0$

$$\mathsf{x}_{t+1} = \mathsf{x}_t - \gamma \nabla f_{i_t}(\mathsf{x}_t), i_t \sim P \textit{ such that } P(i_t = i) = p_i, i = 1, \dots, n$$

achieves linear convergence when $\gamma < \frac{2}{L_{\max}}$*:*

$$\mathbb{E}[||\mathsf{x}_{t+1} - \mathsf{x}^*||^2] \leq \left(1 - \mu(2\gamma - \gamma^2 L_{\max})\right) \mathbb{E}[||\mathsf{x}_t - \mathsf{x}^*||^2],$$

where $L_{\max} = \max_{1 \leq i \leq n} L_i$. *More specially, setting* $\gamma = \frac{1}{L_{\max}}$ *yields the sample complexity* $O(\frac{L_{\max}}{\mu} \log \frac{1}{\epsilon})$.

(c) *(5 points) A natural question one might ask: is it possible to improve the dependence on* $\frac{L_{\max}}{\mu}$ *to* $\frac{\bar{L}}{\mu}$ *with* $\bar{L} = \sum_{i=1}^n p_i L_i$*? Well, this may not be possible for SGD. Consider the special example* $\min_{\mathsf{x}} \frac{1}{n} \sum_{i=1}^n f_i(\mathsf{x})$ *where*

$$\mathsf{x} \in \mathbb{R}^2, f_1(\mathsf{x}) = \frac{n-1}{2}(x_1 - 1)^2, f_2(\mathsf{x}) = \dots = f_n(\mathsf{x}) = \frac{1}{2} x_2^2.$$

Note that in this case $L_{\max} = n - 1, \bar{L} = \frac{2n-1}{n} = O(1)$, $\mu = \frac{n-1}{n}$. *Show that SGD with initial point* $\mathsf{x}_0 = 0$ *and any stepsize requires at least* $\frac{L_{\max}}{\mu}$ *samples in expectation to reach a solution within error less than 1/2, i.e.,* $||\mathsf{x} - \mathsf{x}^*|| \leq 1/2$*?*

(d) *(8 points) Design a modified SGD algorithm for solving (1) that achieves the sample complexity* $O(\frac{\bar{L}}{\mu} \log \frac{1}{\epsilon})$ *with* $\bar{L} = \sum_{i=1}^n p_i L_i$ *under the above assumptions. Please also justify your result.*

Solution:

(a) **To show the second inequality.**

Since F is μ-strongly convex, we have for any $\mathbf{x}, \mathbf{y}$:

$$F(\mathbf{x}) \geq F(\mathbf{y}) + \langle \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \frac{\mu}{2} ||\mathbf{x} - \mathbf{y}||^2,$$

and similarly,

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2} ||\mathbf{x} - \mathbf{y}||^2.$$

Combing these two inequalities, we have

$$\langle \nabla F(\mathbf{x}) - \nabla F(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu ||\mathbf{x} - \mathbf{y}||^2. \tag{2}$$

We reach our conclusion by setting $y = x^*$ and noting that $\nabla F(x^*) = 0$.

To show the first inequality.

Proof option 1: Since f_i is convex and L_i-smooth, we know f_i^* is $\frac{1}{L_i}$-strongly convex. Plugging in $F = f_i^*$ in the above equation with $\mathbf{x} = \nabla f_i(\mathbf{u}), \mathbf{y} = \nabla f(\mathbf{v})$, and invoking the fact that $\mathbf{u} = \nabla f_i^*(\mathbf{x}), \mathbf{v} = \nabla f_i^*(\mathbf{y})$ from Fenchel duality, we have

$$\langle \nabla f_i(\mathbf{u}) - \nabla f_i(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \geq \frac{1}{L_i} ||\nabla f_i(\mathbf{u}) - \nabla f_i(\mathbf{v})||^2. \tag{3}$$

Proof option 2: For ease of notation, let $f = f_i, L = L_i$. Set $\mathbf{z} = \mathbf{y} + \frac{1}{L}(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))$.

$$\begin{aligned}
f(y) - f(x) &= f(y) - f(z) + f(z) - f(x) \\
&\geq -\nabla f(y)^T(z - y) - \frac{L}{2}||y - z||_2^2 + \nabla f(x)^T(z - x) \\
&= \nabla f(x)^T(y - x) - \{\nabla f(x) - \nabla f(y)\}^T(y - z) - \frac{L}{2}||y - z||_2^2 \quad \text{(by plugging in z)} \\
&= \nabla f(x)^T(y - x) + \frac{1}{L}||\nabla f(x) - \nabla f(y)||_2^2 - \frac{1}{2L}||\nabla f(y) - \nabla f(x)||_2^2 \\
&= \nabla f(x)^T(y - x) + \frac{1}{2L}||\nabla f(x) - \nabla f(y)||_2^2
\end{aligned}$$

Exchanging $\mathbf{x}$ and $\mathbf{y}$ and combing the two inequalities imply

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{L} ||\mathbf{x} - \mathbf{y}||^2. \tag{4}$$

(b) First, we have

$$||\mathbf{x}_{t+1} - \mathbf{x}^*||^2 = ||\mathbf{x}_t - \mathbf{x}^*||^2 - 2\gamma \langle \mathbf{x}_t - \mathbf{x}^*, \nabla f_{i_t}(\mathbf{x}_t) \rangle + \gamma^2 ||\nabla f_{i_t}(\mathbf{x}_t)||^2$$

From the first property in (a) and using the fact that $L_i \leq L_{max}$, we have

$$||\mathbf{x}_{t+1} - \mathbf{x}^*||^2 \leq ||\mathbf{x}_t - \mathbf{x}^*||^2 - 2\gamma \langle \mathbf{x}_t - \mathbf{x}^*, \nabla f_{i_t}(\mathbf{x}_t) \rangle + \gamma^2 L_{max} \langle \mathbf{x}_t - \mathbf{x}^*, \nabla f_{i_t}(\mathbf{x}^*) \rangle$$

Taking expectation on both sides,

$$\mathbb{E}[||x_{t+1} - x^*||^2] \leq \mathbb{E}||x_t - x^*||^2 - (2\gamma - \gamma^2 L_{max})\mathbb{E}\langle x_t - x^*, \nabla F(x_t)\rangle$$

Since $\gamma < \frac{2}{L_{max}}$, we have $2\gamma - \gamma^2 L_{max} > 0$. Also, by invoking the strong convexity property from (a), we know $\langle x_t - x^*, \nabla F(x_t)\rangle \geq \mu||x_t - x^*||^2 \geq 0$, and therefore

$$\mathbb{E}[||x_{t+1} - x^*||^2] \leq \mathbb{E}||x_t - x^*||^2 - (2\gamma - \gamma^2 L_{max})\mu\mathbb{E}||x_t - x^*||^2.$$

If we set $\gamma = \frac{1}{L_{max}}$, it leads to

$$\mathbb{E}[||x_t - x^*||^2] \leq (1 - \mu/L_{max})^t \mathbb{E}||x_0 - x^*||^2.$$

(c) In order to reach a solution with small error, we need to examine f_1 at least once. Due to uniform sampling, we need in expectation at least n samples to see f_1. Note that $n = \frac{L_{max}}{\mu}$ in this case.

(d) Consider the weighted SGD algorithm:

$$x_{t+1} = x_t - \frac{1}{L_i}\nabla f_{i_t}(x_t), i_t \sim P \text{ such that } P(i_t = i) = \frac{p_i L_i}{\bar{L}}, i = 1, \ldots, n$$

Note that the above update is equivalent to applying standard SGD with stepsize $\gamma = \frac{1}{\bar{L}}$ on the equivalent problem

$$\min_x \sum_{i=1}^n \tilde{p}_i \tilde{f}_i(x), \text{ with } \tilde{f}_i = \frac{\bar{L}}{L_i} f_i, \tilde{p}_i = \frac{p_i L_i}{\bar{L}}.$$

From (a), we know that the $L_{max}(\tilde{f}_1, \ldots, \tilde{f}_n) = \bar{L}$. This implies the weighted SGD attains the sample complexity to $O(\frac{\bar{L}}{\mu}\log\frac{1}{\epsilon})$.