Continuous Optimization **Spring 2025**

Homework 3: Due 30/05/2025 before 23.55

Lecturer: Aurelien Lucchi

Note: if you use the results from lectures and/or exercise sessions, please state the exact name of a theorem/property you refer to for completeness of your work.

Problem 1 (Newton's method, 10 points):

(a) We consider a function $f\colon \mathbb{R}^2 \to \mathbb{R}$ such that $f(x, y) = x^2 - 2xy + 6y^2$ and apply Newton's method to solve it.

- (i) Derive the expressions for $\nabla f(x, y)$ and $\nabla^2 f(x, y)$. (1 Pt)
- (ii) Perform one step of standard Newton's method starting from $(1, 2)$. (1 Pt)
- (iii) Find all global minima of this function. (1 Pt)
- (iv) Perform one step of standard Newton's method starting from $(2, 4)$. What do you observe? How does the initialization affect the convergence of Newton's method? (1 Pt)

(b) We consider a function $f\colon \mathbb{R}^2 \to \mathbb{R}$ such that $f(x, y) = x^4 + y^4 - 3x^2 - 3y^2$ and apply Newton's method to solve it.

- (i) Derive the expressions for $\nabla f(x, y)$ and $\nabla^2 f(x, y)$. (1 Pt)
- (ii) Find all global minima of this function. (1 Pt)
- (iii) Perform one step of standard Newton's method starting from $(2, 2)$. Does the function value decrease? (1.5 Pts)
- (iv) Perform one step of standard Newton's method starting from $(\frac{1}{2}, \frac{1}{2})$. Does the function value decrease? (1.5 Pts)
- (v) What do you observe? How does the initialization affect the convergence of Newton's method? (1 Pt)

Problem 2 (Proximal Stochastic Gradient Descent, 10 points):

We consider a finite-sum minimization problem with regularization of the form

$$h(\mathbf{x}) \coloneqq \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x})}_{\coloneqq f(\mathbf{x})} + \lambda R(\mathbf{x}), \tag{1}$$

where each individual loss f_i is L_i-smooth. Moreover, we assume that the empirical loss f is μ-strongly convex. To solve this problem, we use proximal stochastic gradient descent (PSGD) of the form

$$\begin{aligned}
&\text{Sample a batch: } S_k \text{ of cardinality } \tau, \text{ i.e., } |S_k| = \tau \\
&\text{Compute a stochastic gradient: } g(\mathbf{x}_k) = \frac{1}{\tau}\sum_{i\in S_k} \nabla f_i(\mathbf{x}_k) \\
&\text{Perform a step of PSGD: } \mathbf{x}_{k+1} = \operatorname{prox}_{\gamma R}(\mathbf{x}_k - \gamma g(\mathbf{x}_k)).
\end{aligned} \tag{2}$$

We define $[n] \coloneqq \{1, \dots, n\}$, $L_{\max} \coloneqq \max\limits_{i\in[n]} L_i$, and $\overline{L} \coloneqq \frac{1}{n}\sum_{i=1}^{n} L_i$. It turns out that the empirical loss f is L-smooth with L satisfying $L \le \overline{L}$.

(a)
- (i) Prove that f is L-smooth where $L \le \overline{L}$. (0.5 Pt)
- (ii) Provide an example of n functions f_i each being L_i-smooth such that $L = L_{\max}$. (0.5 Pt)
- (iii) Provide an example of n functions f_i each being L_i-smooth such that $L \approx \frac{L_{\max}}{n}$. (1 Pt)

The stochastic gradient $g(\mathbf{x})$ of the algorithm satisfies so called *Expected Smoothness* inequality

$$\mathbb{E}[\|g(\mathbf{x}) - \nabla f(\mathbf{x}^*)\|^2] \le 2AD_f(\mathbf{x}, \mathbf{x}^*) + \sigma_*^2,$$

where $D_f(\mathbf{x}, \mathbf{y}) \coloneqq f(\mathbf{x}) - f(\mathbf{y}) - \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$ is a Bregman divergence. Note that the convexity of f implies that $D_f(\mathbf{x}, \mathbf{y}) \ge 0$. Here constants A andσ_*^2 are defined as

$$A \coloneqq \frac{n-\tau}{\tau(n-1)} L_{\max} + \frac{n(\tau-1)}{\tau(n-1)} L, \quad \sigma_*^2 \coloneqq \frac{n-\tau}{\tau(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(\mathbf{x}^*)\|^2 - \|\nabla f(\mathbf{x}^*)\|^2\right).$$

Now we switch to derive the convergence guarantees for PSGD method.

(b) (i) Show that the stochastic gradient $g(\mathbf{x}_k)$ is unbiased estimator of $\nabla f(\mathbf{x})$, i.e., (1 Pt)

$$\mathbb{E}[g(\mathbf{x})] = \nabla f(\mathbf{x}),$$

where the expectation is taken w.r.t. the sampling of the batch S.

(ii) We define a conditional expectation $\mathbb{E}_k[\cdot] \coloneqq \mathbb{E}[\cdot \mid \mathbf{x}_k]$, i.e., w.r.t the σ-algebra defined by $\{\mathbf{x}_0, \ldots, \mathbf{x}_k\}$. In other words, only the randomness of S_k is considered while that of $S_{k-1}, \ldots, S_0$ is frozen. Using properties of the proximity operator, show that the following equality holds (1 Pt)

$$\mathbb{E}_k[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] = \mathbb{E}_k[\|\mathrm{prox}_{\gamma R}(\mathbf{x}_k - \gamma g(\mathbf{x}_k)) - \mathrm{prox}_{\gamma R}(\mathbf{x}^* - \gamma \nabla f(\mathbf{x}^*))\|^2].$$

(iii) Using properties of the proximity operator show that (1 Pt)

$$\mathbb{E}_k[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \leq \mathbb{E}_k[\|\mathbf{x}_k - \gamma g(\mathbf{x}_k) - (\mathbf{x}^* - \gamma \nabla f(\mathbf{x}^*))\|^2].$$

(iv) Using unbiasedness of $g(\mathbf{x}_k)$ show that (2 Pts)

$$\mathbb{E}_k[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \leq \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma\langle \mathbf{x}_k - \mathbf{x}^*, \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*)\rangle + \gamma^2 \mathbb{E}_k[\|g(\mathbf{x}_k) - \nabla f(\mathbf{x}^*)\|^2]$$

(v) Using μ-strong convexity of f, i.e., (1 Pt)

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

and expected smoothness, show that

$$\mathbb{E}_k[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \leq (1 - \gamma\mu)\|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma(1 - \gamma A) D_f(\mathbf{x}_k, \mathbf{x}^*) + \gamma^2 \sigma_*^2.$$

(vi) Using the stepsize restriction $\gamma \leq \frac{1}{A}$, taking full expectation, and using the tower property $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}_k[\cdot]]$, derive that (1 Pt)

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \leq (1 - \gamma\mu)\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|^2] + \gamma^2 \sigma_*^2. \quad (3)$$

(vii) Unrolling (3), show that (1 Pt)

$$\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|^2] \leq (1 - \gamma\mu)^k \|\mathbf{x}_0 - \mathbf{x}^*\|^2 + \frac{\gamma}{\mu}\sigma_*^2.$$

(b) *Bonus:* Assume that the interpolation regime holds, i.e., $\nabla f_i(\mathbf{x}^*) = 0$ for all $i \in [n]$. The iteration complexity of PSGD algorithm is (3 Pts)

$$\text{after } k \geq \max\left\{\frac{2A}{\mu}, \frac{4\sigma_*^2}{\varepsilon\mu^2}\right\} \log\left(\frac{2\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{\varepsilon}\right) \text{ iterations we have } \mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|^2] \leq \varepsilon.$$

Assume that computing one stochastic gradient $\nabla f_i(\mathbf{x})$ for any $i \in [n]$ costs 1 time unit. This implies that computing mini-batch stochastic gradient $g(\mathbf{x})$ costs τ. Show that vanilla gradient descent, i.e., the case when $\tau = n$, achieves the fastest convergence in this regime w.r.t. time.

Problem 3 (Stochastic Coordinate Descent, 10 points):

We consider a problem of minimizing

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}), \quad (4)$$

where f is μ-strongly convex, i.e., for all $\mathbf{y}, \mathbf{x} \in \mathbb{R}^d$ we have

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

Moreover, we assume that f is $\mathbf{L}$-smooth, where $\mathbf{L} = \mathrm{diag}(L_1, \ldots, L_d)$, $L_i > 0$ for all $i \in [d]$, namely for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ we have

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{1}{2}\langle \mathbf{L}(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x}\rangle = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_{\mathbf{L}}^2. \quad (5)$$

Here for any diagonal matrix $\mathbf{W}$ the $\mathbf{W}$-norm of a vector $\mathbf{a}$ means (above we use this notation for $\mathbf{W} = \mathbf{L}$)

$$\|\mathbf{a}\|_{\mathbf{W}}^2 \coloneqq \mathbf{a}^\top \mathbf{W}\mathbf{a}.$$

Let $p_i > 0, i \in [d]$ be a discrete probability distribution, i.e., $\sum_{i=1}^d p_i = 1$. Let a matrix $\mathbf{P} = \operatorname{diag}(1/p_1, \ldots, 1/p_d)$ and $\mathbf{C}_k = \operatorname{diag}(c_1^k, \ldots, c_d^k)$, where

$$c_i^k = \begin{cases} 1/p_i, & \text{with probability } p_i \\ 0, & \text{otherwise} \end{cases}.$$

We assume that $\{c_i^k\}_{i=1}^d$ are independent random variables. To solve (4), we consider the Coordinate Descent (CD) method of the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma \mathbf{C}_k \nabla f(\mathbf{x}_k).$$

This algorithm can be rewritten in a simpler form as follows

$$[\mathbf{x}_{k+1}]_i = \begin{cases} [\mathbf{x}_k]_i - \gamma \frac{[\nabla f(\mathbf{x}_k)]_i}{p_i}, & \text{with probability } p_i \\ [\mathbf{x}_k]_i, & \text{otherwise} \end{cases},$$

where $[\cdot]_i$ denotes the i-th entry of the vector. The goal of this problem is to show the convergence of the CD algorithm. Let $\mathbb{E}_k[\cdot]$ be a conditional expectation w.r.t. a σ-algebra generated by $\{\mathbf{x}_0, \ldots, \mathbf{x}_k\}$. In other words, only the randomness of $\mathbf{C}_k$ is considered while that of $\mathbf{C}_{k-1}, \ldots, \mathbf{C}_0$ is frozen.

(a) Show that the stochastic gradient $\mathbf{C}_k \nabla f(\mathbf{x}_k)$ is unbiased, i.e., $\mathbb{E}_k[\mathbf{C}_k \nabla f(\mathbf{x}_k)] = \nabla f(\mathbf{x}_k)$. (1 Pt)

(b) Using the definition of $\mathbf{C}_k$ show that (1 Pt)

$$\mathbb{E}_k[\|\mathbf{C}_k \nabla f(\mathbf{x}_k)\|^2] = \nabla f(\mathbf{x}_k)^\top \mathbf{P} \nabla f(\mathbf{x}_k) = \|\nabla f(\mathbf{x}_k)\|_{\mathbf{P}}^2.$$

(c) Let us define $L_P \coloneqq \max_{i \in [d]} \frac{L_i}{p_i}$. This implies $\mathbf{P}^{1/2}\mathbf{L}\mathbf{P}^{1/2} \le L_P \mathbf{I}$. Show that (1 Pt)

$$\mathbf{PLP} \le L_P \mathbf{P}.$$

(d) Using $\mathbf{L}$-smoothness inequality (5) with $\mathbf{y} = \mathbf{x} - \alpha \mathbf{P} \nabla f(\mathbf{x}), \alpha = \frac{1}{L_P}$, show that (2 Pts)

$$f(\mathbf{y}) \le f(\mathbf{x}) - \frac{\alpha}{2}\|\nabla f(\mathbf{x})\|_{\mathbf{P}}^2.$$

(e) Using the previous result show that (1 Pt)

$$\|\nabla f(\mathbf{x})\|_{\mathbf{P}}^2 \le 2L_P(f(\mathbf{x}) - f(\mathbf{x}^*)).$$

(f) Using previous results show that (1 Pt)

$$\mathbb{E}_k[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \le \|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma\langle \mathbf{x}_k - \mathbf{x}^*, \nabla f(\mathbf{x}_k)\rangle + 2L_P\gamma^2(f(\mathbf{x}_k) - f(\mathbf{x}^*)).$$

(g) Using μ-strong convexity, show that (1 Pt)

$$\mathbb{E}_k[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \le (1 - \gamma\mu)\|\mathbf{x}_k - \mathbf{x}^*\|^2 - 2\gamma(1 - L_P\gamma)(f(\mathbf{x}_k) - f(\mathbf{x}^*)).$$

(h) Takin full expectation, using tower property $\mathbb{E}[\cdot] = \mathbb{E}[\mathbb{E}_k[\cdot]]$, and stepsize restriction $\gamma \le 1/L_P$, show that (1 Pt)

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2] \le (1 - \gamma\mu)\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|^2].$$

(i) Unrolling the recursion, show that (1 Pt)

$$\mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|^2] \le (1 - \gamma\mu)^k\|\mathbf{x}_0 - \mathbf{x}^*\|^2.$$

Bonus: The iteration complexity of CD with a stepsize $\gamma = \frac{1}{L_P}$ is (2 Pts)

$$\text{after } k \ge \frac{L_P}{\mu}\log\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{\varepsilon} \text{ iterations we have } \mathbb{E}[\|\mathbf{x}_k - \mathbf{x}^*\|^2] \le \varepsilon.$$

This implies that the smaller the ratio $\frac{L_P}{\mu}$ is, the faster the convergence is. Let us consider two sampling strategies $\hat{p}_i = \frac{1}{n}$ for all $i \in [d]$, and $\tilde{p}_i = \frac{L_i}{\sum_{i=1}^d L_i}$ for all $i \in [d]$. Which strategies leads to faster convergence?