RWTH Aachen University **Winter 2023/24**

Instructor: Prof. Dr. Semih Çaycı
Teaching Assistant: Johannes Müller, M.Sc.

Mathematical Foundations of Deep Learning (11.80020)
Assignment 2

Due: Tuesday, Nov. 21st, till 2pm as PDF via Moodle upload, TeX submission are encouraged
Each problem is worth 4 points, there are 20 points on this sheet. Submission in pairs is possible.

Q1. (Convexity of empirical risk) Consider a linear model, i.e., $f_\theta(x) = \theta^\top \Phi(x)$ for a fixed feature function $\Phi\colon \mathbb{X} \to \mathbb{R}^p$, where $\theta \in \Theta$ for a convex parameter set $\Theta \subseteq \mathbb{R}^p$. Show that the empirical risk $\hat{\mathcal{R}}_S\colon \Theta \to \mathbb{R}$ is convex for the following sample losses $\ell\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$:

(a) *l^2-loss:* $\ell(\hat{y}, y) \coloneqq \frac{1}{2}(\hat{y} - y)^2$.

(b) *l^2-loss proxy for $0-1$-loss:* $\ell(\hat{y}, y) \coloneqq (1 - \hat{y}y)^2$.

(c) *Logistic loss:* $\ell_{log}(\hat{y}, y) \coloneqq \log(1 + e^{-\hat{y}y})$.

(d) *Hinge loss:* $\ell_{Hinge}(\hat{y}, y) \coloneqq \max\{1 - \hat{y}y, 0\}$.

Further, construct an example of a linear model such that the empirical risk is non-convex for the $0-1$-loss $\ell_{0-1}(\hat{y}, y) = \mathbb{1}_{\hat{y}y \leq 0}$.

Solution: Note that in general sums of convex functions are convex as well as the composition of a linear function with a convex function. Therefore, in order to show the convexity of

$$\theta \mapsto \sum_{i=1}^n \ell(\theta^\top \Phi(x_i), y_i)$$

it suffices to show the convexity of $\ell(\cdot, y)\colon \mathbb{R} \to \mathbb{R}$, which we do for the individual losses:

(a) It holds that $\partial_{\hat{y}}^2 \ell(\cdot, y) = 2 > 0$ and hence $\ell(\cdot, y)$ is convex by the second order convexity condition, in fact 2-strongly convex.

(b) It holds that $\partial_{\hat{y}}^2 \ell(\cdot, y) = 2y^2 \geq 0$.

(c) It holds that

$$\partial_{\hat{y}}^2 \ell(\cdot, y) = \frac{y^2 e^{-\hat{y}y}}{(1 + e^{-\hat{y}y})^2} \geq 0.$$

(d) Finally, note that $\hat{y} \mapsto 1 - \hat{y}y$ and $\hat{y} \mapsto 0$ are linear hence convex and that the maximum of convex functions is again convex.

Take now the case $\mathbb{X} = \mathbb{Y} = \mathbb{R}$ with $f_\theta(x) = \theta$ and the training set $S = \{(0, 1)\}$. Then the resulting empirical loss is given by

$$\mathcal{R}(\theta) = \mathbb{1}_{\theta \leq 0},$$

which is non convex.

Q2. (Preconditioned GD) Consider a differentiable function $g\colon \mathbb{R}^d \to \mathbb{R}$ and fix a symmetric positive definite matrix $A \in \mathbb{R}^{d\times d}$ as well as $\eta > 0$. For $\theta_0 \in \mathbb{R}^d$ we choose

$$\theta_1 \in \arg\min_{\theta\in\mathbb{R}^d} \left\{ g(\theta_0) + \nabla g(\theta_0)^\top(\theta - \theta_0) + \frac{1}{2\eta}\|\theta_0 - \theta\|_A^2 \right\}, \tag{1}$$

where $\|\theta\|_A^2 \coloneqq \theta^\top A\theta$ denotes the norm induced by A. Show that there is a unique minimum θ_1 and that

$$\theta_1 = \theta_0 - \eta A^{-1}\nabla g(\theta_0). \tag{2}$$

Remark: In particular, choosing $A = I$ recovers the vanilla gradient descent update and $A = \nabla^2 g(\theta_0)$ (if existent) recovers Newton's method. The update rule (1) is a specific example of the *mirror descent* with Bregman divergence $D(\theta,\phi) = \frac{1}{2}\|\theta - \phi\|_A^2$ and (2) is called *preconditioned gradient descent*.

Solution: Note that the objective

$$f(\theta) = g(\theta_0) + \nabla g(\theta_0)^\top(\theta - \theta_0) + \frac{1}{2\eta}\|\theta_0 - \theta\|_A^2$$

if $\eta\lambda_{min}(A)$-strongly convex, where $\lambda_{min}(A) > 0$ denotes the smallest eigenvalue of A. In particular, a unique optimizer θ_1 of f exists. The optimizer is uniquely characterized by the stationarity condition

$$0 = \nabla f(\theta_1) = \nabla g(\theta_0) + \frac{1}{\eta}A(\theta_0 - \theta_1).$$

Solving for θ_1 yields (2).

Q3. (High-probability bounds for projected SGD) Consider a differentiable convex function $g\colon \mathbb{R}^d \to \mathbb{R}$, an $\mathbb{R}^d$-valued random variable θ_0 as well as the projected stochastic gradient update rule

$$\widetilde{\theta}_{t+1} = \theta_t - \eta u_t, \tag{3}$$

$$\theta_{t+1} = \Pi_{B_2(0,R)}\widetilde{\theta}_{t+1}, \tag{4}$$

where $\Pi_{B_2(0,R)}$ denotes the Euclidean projection onto the closed Euclidean ball $B_2(0,R) = \{x \in \mathbb{R} : \|x\|_2 \le R\}$ with radius $R > 0$. Assume that $(u_t)_{t\in\mathbb{N}}$ is a sequence of $\mathbb{R}^d$-valued random variables satisfying $\|u_t\|_2 \le L$ almost surely that are unbiased gradient estimators, i.e., $\mathbb{E}[u_t|\mathcal{F}_t] = \nabla g(\theta_t)$, where $\mathcal{F}_t = \sigma(\theta_0,\dots,\theta_t)$. Show that for the step size $\eta = \frac{\sqrt{2}R}{\sqrt{T}L}$ it holds with probability at least $1-\delta$ that

$$g\left(\frac{1}{T}\sum_{t=0}^{T-1}\theta_t\right) - g^\star \le \frac{\sqrt{2}RL}{\sqrt{T}}\left(1 + \sqrt{\log(2/\delta)}\right). \tag{5}$$

Hint: Define $D_t \coloneqq \mathbb{E}[u_t^\top(\theta_t - \theta^\star)|\mathcal{F}_t] - u_t^\top(\theta_t - \theta^\star)$. Is this a martingale difference sequence? Use an appropriate concentration inequality.

Solution: Note that a minimizer $\theta^\star$ exists since we optimize a continuous function over a closed ball, which is compact. Just like in the deterministic case we consider the Lyapunov function $\mathcal{L}(\theta) := \|\theta - \theta^*\|_2^2$. Using that the projection is a non-expansive operator we estimate

$$\begin{aligned}
\|\theta_{t+1} - \theta^\star\|_2^2 &\leq \|\widetilde{\theta}_{t+1} - \theta^\star\|_2^2 \\
&= \|\theta_t - \eta u_t - \theta^\star\|_2^2 \\
&= \|\theta_t - \theta^\star\|_2^2 - 2\eta u_t^\top(\theta_t - \theta^\star) + \eta^2\|u_t\|_2^2 \\
&\leq \|\theta_t - \theta^\star\|_2^2 - 2\eta\mathbb{E}[u_t^\top(\theta_t - \theta^\star)|\mathcal{F}_t] + 2\eta D_t + \eta^2 L^2.
\end{aligned}$$

Now, note that due to the convexity and since u_t is an unbiased gradient estimator we obtain

$$\mathbb{E}[u_t^\top(\theta_t - \theta^\star)|\mathcal{F}_t] = \nabla g(\theta_t)^\top(\theta_t - \theta^\star) \leq g(\theta_t) - g(\theta^\star)$$

and hence

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) \leq -2\eta c(\theta_t) + 2\eta D_t + \eta^2 L^2$$

with the cost $c(\theta) = g(\theta) - g(\theta^\star)$. Using $\|\theta_0\|_2^2 \leq R^2$ the Lyapunov-drift inequality yields

$$g\left(\frac{1}{T}\sum_{t=0}^{T-1}\theta_t\right) - g(\theta^\star) \leq \frac{R^2}{\eta T} + \frac{\eta L^2}{2} + \frac{1}{T}\sum_{t=0}^{T-1} D_t.$$

Note that D_t is a martingale difference sequence and bounded by $2L$ and hence the Azuma-Hoeffding ineuqality yields

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=0}^{T-1} D_t\right| > 2RL\sqrt{\frac{\log(2/\delta)}{2T}}\right) \leq \delta.$$

Therefore, when choosing the step size $\eta = \frac{\sqrt{2}R}{\sqrt{T}L}$ it holds with probability at least $1-\delta$ that

$$g\left(\frac{1}{T}\sum_{t=0}^{T-1}\theta_t\right) - g^\star \leq \frac{\sqrt{2}RL}{\sqrt{T}}\left(1 + \sqrt{\log(2/\delta)}\right).$$

Q4. (Parameter convergence of GD and SGD) Let $g : \mathbb{R}^d \to \mathbb{R}$ be an α-strongly-convex and β-smooth function with $\mathrm{dom}(g) = \mathbb{R}^d$, and unique optimal point $\theta^\star \in \mathbb{R}^d$.

(a) Consider gradient descent with constant step-size: $\theta_{t+1} = \theta - \eta\nabla g(\theta_t)$ with an arbitrary initial point $\theta_0 \in \mathbb{R}^d$. Then, show that, with the step-size choice $\eta = \frac{\alpha}{2\beta^2}$, the following bound is achieved:

$$\|\theta_t - \theta^\star\|_2^2 \leq \left(1 - \frac{\alpha^2}{4\beta^2}\right)^t \|\theta_0 - \theta^\star\|_2^2,$$

for any $t \geq 1$.

Hint: Use the potential function $\mathcal{L}(\theta) = \|\theta - \theta^\star\|_2^2$. Also, use the fact that $\nabla g(\theta^\star) = 0$.

Solution: In order to use the Lyapunov drift theory, we aim to bound the Lyapunov drift $\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t)$. By the α-strong convexity and β-smoothness we have

$$\begin{aligned}
\|\theta_{t+1} - \theta^\star\|_2^2 &= \|\theta_t - \theta^\star - \eta\nabla g(\theta_t)\|_2^2 \\
&= \|\theta_t - \theta^\star\|_2^2 + 2\eta\nabla g(\theta_t)^\top(\theta^\star - \theta_t) + \eta^2\|\nabla g(\theta_t)\|_2^2 \\
&\leq \|\theta_t - \theta^\star\|_2^2 + (g(\theta^\star) - g(\theta_t)) - \eta\alpha\|\theta_t - \theta^\star\|_2^2 + \eta^2\beta^2\|\theta_t - \theta^\star\|_2^2 \\
&\leq (1 - \eta\alpha + \eta^2\beta^2)\|\theta_t - \theta^\star\|_2^2 \\
&= \left(1 - \frac{\alpha^2}{4\beta^2}\right)^t \|\theta_t - \theta^\star\|_2^2.
\end{aligned} \tag{6}$$

Iterating over t now yields the claim.

(b) Consider stochastic gradient descent with constant step-size: $\theta_{t+1} = \theta_t - \eta u_t$ for a sequence $(u_t)_{t\geq 0}$ of $\mathbb{R}^d$-valued random variables with:

$$\mathbb{E}[u_t|\mathcal{F}_t] = \nabla g(\theta_t) \quad \text{and } \mathbb{E}\left[\left\|u_t - \nabla g(\theta_t)\right\|_2^2 |\mathcal{F}_t\right] \leq \nu^2$$

almost surely for all $t \geq 0$, where $\mathcal{F}_t = \sigma(\theta_0, \dots, \theta_t)$. Show that, for $\eta > 0$ sufficiently small such that $\rho = \eta(\alpha - 2\eta\beta^2) \in (0,1)$ it holds that

$$\mathbb{E}\left[\|\theta_t - \theta^\star\|_2^2\right] \leq (1-\rho)^t \|\theta_0 - \theta^\star\|_2^2 + \frac{2\eta^2\nu^2}{\rho},$$

for any $t \geq 1$.

Hint: The following inequality can be useful: $\|\theta + \phi\|_2^2 \leq 2\|\theta\|_2^2 + 2\|\phi\|_2^2$ for any $\theta, \phi \in \mathbb{R}^d$.

Solution: We compute

$$\begin{aligned}\|\theta_{t+1} - \theta^\star\|_2^2 &= \|\theta_t - \theta^\star - \eta\nabla g(\theta_t) + (\nabla g(\theta_t) - u_t)\|_2^2 \\ &= \|\theta_t - \theta^\star - \eta\nabla g(\theta_t)\|_2^2 + 2\eta(\theta_t - \theta^\star - \eta\nabla g(\theta_t))^\top(\nabla g(\theta_t) - u_t) \qquad (7)\\ &\quad + \eta^2\|\nabla g(\theta_t) - u_t\|_2^2.\end{aligned}$$

Since θ_t is $\mathcal{F}_t$ measurable and $\mathbb{E}[u_t|\mathcal{F}_t] = \nabla g(\theta_t)$ we have

$$\mathbb{E}\left[(\theta_t - \theta^\star - \eta\nabla g(\theta_t))^\top(\nabla g(\theta_t) - u_t)|\mathcal{F}_t\right] = (\theta_t - \theta^\star - \eta\nabla g(\theta_t))^\top \mathbb{E}\left[\nabla g(\theta_t) - u_t|\mathcal{F}_t\right] = 0.$$

Hence, taking the conditional expectation with respect to $\mathcal{F}_t$ to (7) and using the estimate (6) we obtain

$$\begin{aligned}\mathbb{E}\left[\|\theta_{t+1} - \theta^\star\|_2^2|\mathcal{F}_t\right] &= \|\theta_t - \theta^\star - \eta\nabla g(\theta_t)\|_2^2 + \eta^2\mathbb{E}\left[\|\nabla g(\theta_t) - u_t\|_2^2|\mathcal{F}_t\right] \\ &\leq (1 - \eta\alpha + \eta^2\beta^2)\|\theta_t - \theta^\star\|_2^2 + \eta^2\nu^2.\end{aligned}$$

Taking the expectation and using the tower property we obtain

$$\mathbb{E}\left[\|\theta_{t+1} - \theta^\star\|_2^2\right] \leq (1 - \eta\alpha + \eta^2\beta^2)\mathbb{E}\left[\|\theta_t - \theta^\star\|_2^2\right] + \eta^2\nu^2.$$

With $\delta \coloneqq \eta(\alpha - \eta\beta^2) \in (0,1)$, iterating over t and using $\sum_{t=0}^T (1-\delta)^t \leq \sum_{t\in\mathbb{N}}(1-\delta)^t = \delta^{-1}$ we obtain

$$\mathbb{E}\left[\|\theta_t - \theta^\star\|_2^2\right] \leq (1-\delta)^t\|\theta_0 - \theta^\star\|_2^2 + \frac{\eta^2\nu^2}{\delta}.$$

Note that $\delta > \rho$.

Remark: In the lecture, we used a different potential function $\theta \mapsto g(\theta) - g(\theta^\star)$ along with the descent lemma and Polyak-Lojasiewicz inequality to prove the exponential convergence rate for *function values* under GD. In this question, we use a different potential function to prove exponential convergence rate for *parameters*. Notice that, in the case of SGD, one can use an arbitrarily small step-size $\eta > 0$ to mitigate the impact of ν^2, but this would lead $(1-\rho)$ to be closer to 1.

Q5. (Linear l^2-regression: The underparametrized case) Consider a linear model, i.e., $f_\theta(x) = \theta^\top\Phi(x)$ for a fixed feature function $\Phi\colon \mathbb{X} \to \mathbb{R}^p$, where $\theta \in \mathbb{R}^p$. Further, we consider the l^2 sample loss $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$. In particular, **Q1** implies that the empirical risk

$$L(\theta) = \hat{\mathcal{R}}_S(f_\theta) = \frac{1}{2n}\sum_{i=1}^n \left(\theta^\top\Phi(x_i) - y_i\right)^2 = \frac{1}{2n}\|\Phi(X)\theta - Y\|_2^2,$$

where $\Phi(X)_{ij} \coloneqq \Phi(x_i)_j$ and $Y_i = y_i$ is convex.

(a) *Normal equation:* Compute the gradient $\nabla L(\theta)$ and characterize the set of minimizers

$$\left\{\theta \in \mathbb{R}^p : L(\theta) = \inf_{\theta'} L(\theta')\right\}.$$

Solution: The gradient is given by

$$\nabla L(\theta) = \Phi(X)^\top(\Phi(X)\theta - Y) = \Phi(X)^\top\Phi(X)\theta - \Phi(X)^\top Y. \tag{8}$$

Since the objective is convex, global optima are characterised by the stationarity condition $\nabla L(\theta) = 0$ which is called *normal equation* and takes the form

$$\Phi(X)^\top\Phi(X)\theta = \Phi(X)^\top Y. \tag{9}$$

(b) *Lipschitz gradients:* Consider the Gramian matrix $G = \Phi(X)^\top\Phi(X) \in \mathbb{R}^{p\times p}$ with entries

$$G_{ij} = \sum_{k=1}^n \Phi(x_k)_i\Phi(x_k)_j.$$

Show $\nabla^2 L = G$ and conclude that L has $\lambda_{max}(G)$-Lipschitz gradients, where $\lambda_{max}(G)$ denotes the largest eigenvalue of G.

Solution: From (8) we find that $\nabla^2 L(\theta) = \Phi(X)^\top\Phi(X) = G$. To see the smoothness condition, we estimate

$$\|\nabla L(\theta) - \nabla L(\theta')\|_2 = \|G(\theta-\theta')\|_2 \le \lambda_{max}(G)\cdot\|\theta-\theta'\|_2.$$

(c) *Strong convexity:* Show that if G is full rank then the objective L is $\lambda_{min}(G)$-strongly convex, where $\lambda_{min}(G)$ denotes the smallest eigenvalue of G. In paricular, this implies the existence of a uniquene of minimizer $\theta^\star$ of L. *Remark:* Note that G has full rank if and only if $n \ge p$ and $\Phi(X)$ has rank p. We refer to problems with $n \ge p$ as *underparametrized.*

Solution: Since $\nabla^2 L = G$ the quadratic function L is $\lambda_{min}(G)$-strongly convex.

(d) *Linear convergence of GD:* Show that if G has full rank the GD iterates

$$\theta_{t+1} = \theta_t + \eta\nabla L(\theta)$$

with step size $\eta = \lambda_{max}(G)^{-1}$ satisfy

$$L(\theta_T) - L(\theta^\star) \le e^{-\frac{T}{\kappa(G)}}\cdot(L(\theta_0) - L(\theta^\star)),$$

where $\kappa(G) = \frac{\lambda_{max}(G)}{\lambda_{min}(G)}$ is the condition number of G.

Solution: This is a direct consequence of the linear convergence result of gradient descent for strongly convex functions with Lipschitz gradients.

(e) *Tikhonov regularization / weight decay:* For $\lambda \ge 0$ we set $L_\lambda(\theta) := L(\theta) + \frac{\lambda}{2}\|\theta\|_2^2$. Show that the gradient descent updates $\theta_{t+1} = \theta_t + \eta_t\nabla L_\lambda(\theta_t)$ satisfy

$$\theta_{t+1} = (1-\lambda\eta_t)\theta_t - \eta_t\nabla L(\theta_t).$$

Further, L_λ is $(\lambda_{min}(G)+\lambda)$-strongly convex, in particular, for $\lambda > 0$ there is a unique minimizer $\theta^\star_\lambda$ of L_λ. Further, show that the gradient descent updates satisfy

$$L_\lambda(\theta_T) - L_\lambda(\theta^\star_\lambda) \le e^{-\frac{\lambda_{min}(G)+\lambda}{\lambda_{max}(G)+\lambda}\cdot T}\cdot(L_\lambda(\theta_0) - L_\lambda(\theta^\star_\lambda)) \quad \text{for } T \in \mathbb{N}.$$

Remark: Note that we made the problem strongly convex at the expense of changing the objective and therefore introducing an *regularization bias*.

Solution: Note that L_λ is a again a quadratic function with $\nabla^2 L_\lambda = G + I$. This yields the $(\lambda_{min}(G) + \lambda)$-strong convexity and $(\lambda_{max}(G) + \lambda)$-Lipschitz gradients. Again, the linear convergence result of gradient descent for strongly convex functions with Lipschitz gradients yields the result.

(f) **Bonus (1 point):** *Faster convergence for alternative loss:* Assume that $n = p$ and that $\Phi(X)$ has full rank and is symmetric and positive definite. Consider the alternative loss

$$g(\theta) := \frac{1}{2}\|\theta\|^2_{\Phi(X)} - \theta^\top Y,$$

where $\|\theta\|_{\Phi(X)} = \theta^\top \Phi(X)\theta$. Show that g has the same unique minimizer $\theta^\star$ as L, has $\lambda_{max}(\Phi(X))$-Lipschitz gradients and is $\lambda_{min}(\Phi(X))$-strongly convex. Conclude that GD iterates $(\theta_t)_{t\in\mathbb{N}}$ of g with step size $\eta = \lambda_{min}(\Phi(X))^{-1}$ satisfy

$$L(\theta_T) - L(\theta^\star) \le e^{-\frac{T}{\kappa(\Phi(X))}} \cdot (L(\theta_0) - L(\theta^\star))$$

and show that $\kappa(G) = \kappa(\Phi(X))^2 \ge \kappa(\Phi(X))$ with equality if and only if $\Phi(X) = \alpha I$ for some $\alpha > 0$.

Remark: Note that the proximal loss in (1) uses this alternative formulation in order to compute the preconditioned gradient $w = A^{-1}\nabla g(\theta)$. Part (e) shows that running gradient descent on this objective converges faster than running gradient descent on the objective $w \mapsto \frac{1}{2}\|Aw - \nabla g(\theta)\|^2$.

Solution: Again g is quadratic with $\nabla^2 g = \Phi(X)$. Now the same argument as before yields the convergence result. Note that $\kappa(G) = \kappa(\Phi(X)^\top\Phi(X)) = \kappa(\Phi(X))^2$ as the spectrum of $\Phi(X)^\top\Phi(X)$ consists of the squared eigenvalues of $\Phi(X)$. Finally, note that $\kappa(\Phi(X)) \ge 1$ and $\kappa(\Phi(X)) = 1$ if and only if $\Phi(X) = \alpha I$ for some α.

(g) **Bonus (1 point):** *One step convergence of Newton's method:* Assume that G has full rank. Show that Newton's method with step size $\eta = 1$ converges in one iteration, i.e.,

$$\theta_0 - \nabla^2 L(\theta_0)^{-1}\nabla L(\theta_0) = \theta_0 - G^{-1}\nabla L(\theta_0) = \theta^\star \quad \text{for all } \theta_0 \in \mathbb{R}^p.$$

Where is the caveat?

Solution: It suffices to show that $\theta_1 = \theta_0 - G^{-1}\nabla L(\theta_0)$ satisfies the normal equation (9). Using (8) we find that

$$G\theta_1 = G(\theta_0 - G^{-1}\nabla L(\theta_0)) = G\theta_0 - \nabla L(\theta_0) = G\theta_0 - G\theta_0 + \Phi(X)^\top Y = \Phi(X)^\top Y.$$

Note: The following are bonus problems worth 4 points per problem.

Q6. (Bonus problem: Gradient descent for nonconvex and smooth functions) Let $g\colon \mathbb{R}^d \to \mathbb{R}$ be a β-smooth function (not necessarily convex). Then, for any $\theta_0 \in \mathrm{dom}(g)$, prove that gradient descent with step-size $\eta = 1/\beta$ yields the following:

$$\min_{t=0,1,\dots,T-1} \|\nabla g(\theta_t)\|_2^2 \le \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla g(\theta_t)\|_2^2 \le \frac{2\beta}{T}\Big(g(\theta_0) - g(\theta^\star)\Big), \tag{10}$$

for any $T \ge 1$.

Hint: The descent lemma for β-smooth functions can be useful.