Optimization in Machine Learning | **Universität Mannheim**
FSS 2023 | Prof. Simon Weißmann, Felix Benning

Sheet 5

For the exercise class on the 11.05.2023.
Hand in your solutions by 12:00 in the exercise on Thursday 11.05.2023.

Exercise 1 (Conditional Expectation). **(2 Points)**
Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{F}$ a subalgebra of $\mathcal{A}$ and X, Y random vectors. Prove for $\mathcal{F}$-measurable $X \in \mathbb{R}^d$ that we have

$$\mathbb{E}[\langle X, Y \rangle \mid \mathcal{F}] = \langle X, \mathbb{E}[Y \mid \mathcal{F}] \rangle$$

Exercise 2 (Convexity and Expectation). **(2 Points)**
Let Z be a random variable. Let $f(x) := f(x, Z)$ be a random function ($f(x, \omega) = f(x, Z(\omega))$ if you want) and its expectation

$$F(x) = \mathbb{E}[f(x)]$$

Is f almost surely convex if and only if F is convex? Prove or disprove both directions.

Exercise 3 (Convergence of SGD on Strongly Convex Functions). **(2 Points)**
In the lecture we proved for L-smooth functions F and X_n generated by Algorithm 6 (SGD)

$$\|\nabla F(X_n)\|^2 \to 0 \quad \text{a.s.}$$

If we additionally have strong convexity of F, prove $\|X_n - x_*\| \to 0$ almost surely.

Exercise 4 (Swap Integration with Differentiation). **(9 Points)**

(i) What formal requirements on $f : V \times \Omega \to \mathbb{R}$ with $U \subseteq \mathbb{R}$ and measure μ on Ω are needed, for the following argument using the fundamental theorem of calculus (FTC) to work?

$$\begin{aligned}
\frac{\partial}{\partial t} \int_\Omega f(t_0, \omega) d\mu(\omega) &\overset{\text{linear}}{=} \lim_{\epsilon \to 0} \int \frac{f(t_0 + \epsilon, \omega) - f(t_0, \omega)}{\epsilon} d\mu(\omega) \\
&\overset{\text{FTC II}}{=} \lim_{\epsilon \to 0} \int \frac{1}{\epsilon} \int_{t_0}^{t_0+\epsilon} \frac{\partial}{\partial t} f(t, \omega) dt d\mu(\omega) \\
&\overset{\text{Fubini}}{=} \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int_{t_0}^{t_0+\epsilon} \int \frac{\partial}{\partial t} f(t, \omega) d\mu(\omega) dt \\
&\overset{\text{def.+lin.}}{=} \frac{d}{dy} \int_{t_0}^{y} \int \frac{\partial}{\partial t} f(t, \omega) d\mu(\omega) dt \Big|_{y=t_0} \\
&\overset{\text{FTC I}}{=} \int \frac{\partial}{\partial t} f(t, \omega) d\mu(\omega) dt.
\end{aligned}$$

Formulate the corresponding theorem. (5 pts)

(ii) We want to find an example for a function, where you can not swap integration with differentiation. So for a function $f(t, \omega)$ we need some t_0 such that

$$\frac{\partial}{\partial t} \int_\Omega f(t_0, \omega) d\omega \neq \int_\Omega \frac{\partial}{\partial t} f(t_0, \omega) d\omega.$$

For this consider $f(t, \omega) = t^3 e^{-t^2 \omega}$. Prove the inequality at $t_0 = 0$ and $\Omega = [0, \infty)$. Why is this not a contradiction to (i)? (4 pts)

Hint. *It is helpful to calculate the entire function*

$$t \mapsto \int_0^\infty \frac{\partial}{\partial t} f(t, \omega) d\omega.$$

Exercise 5 (SGD on quadratic functions)**.** **(9 Points)**

Throughout we use the notation for SGD

$$X_{n+1} = X_n - \alpha_n \nabla f_{n+1}(X_n)$$

using $X_0 = x_0 \in \mathbb{R}^d$ with sample errors $\epsilon_n = f_n(x) - F(x)$ for stochastic gradients

$$f_n(x) := f(x, Z_n) = F(x) + \epsilon_n$$

for sample data $(Z_n)_n$ with $Z_n \overset{\text{iid}}{\sim} \mu$ random vectors in $\mathbb{R}^d$. Additionally we write for GD

$$x_{n+1} = x_n - \alpha_n F(x_n).$$

(i) Prove for any y_0 and the recursion

$$y_{n+1} := y_n - \tfrac{1}{n+1}(y_n - z_{n+1})$$

that y_n is a running mean

$$y_n = \frac{1}{n} \sum_{k=1}^{n} z_k =: \bar{z}_n, \quad \forall n \in \mathbb{N} \qquad \text{(1 pt)}$$

(ii) Let $Z \in \mathbb{R}^d$ be a random vector and consider the sample loss

$$f(x, Z) := \tfrac{1}{2}\|x - Z\|_H^2 \overset{\text{recall}}{=} \tfrac{1}{2}\langle x - Z, H(x - Z)\rangle$$

Prove that

$$F(x) = \mathbb{E}[f(x, Z)] = \tfrac{1}{2}\|x - x_*\|_H^2 + \text{const.}$$

with

$$x_* = \operatorname*{argmin}_x \mathbb{E}[\|x - Z\|_H^2].$$

What is x_*? What is the (in the L^2 sense) optimal step size for SGD in the case $H = \mathbb{I}$? (4 pts)

Hint. *The mean is the minimum variance estimator for the expectation.*

Proof. We have

$$
\begin{aligned}
2\mathbb{E}[f(x,Z)] = \mathbb{E}\|x-Z\|_H^2 &= \|x\|_H^2 - 2\mathbb{E}\langle x, Z\rangle_H + \mathbb{E}\|Z\|_H^2 \\
&= \|x\|_H^2 - 2x^T H\mathbb{E}[Z] + \|\mathbb{E}[Z]\|_H^2 + (\mathbb{E}\|Z\|_H^2 - \|\mathbb{E}[Z]\|_H^2) \\
&= \|x - \mathbb{E}[Z]\|_H^2 + \underbrace{(\mathbb{E}\|Z\|_H^2 - \|\mathbb{E}[Z]\|_H^2)}_{=\text{const.}}.
\end{aligned}
$$

So $x_* = \mathbb{E}[Z]$ does the job. We know that $\bar{Z}_n$ is the minimum variance estimator of $\mathbb{E}[Z]$, so it would be perfect if $X_n = \bar{Z}_n$. And with $\alpha_n = \frac{1}{n+1}$ this is in fact possible if we recall

$$\nabla f_n(x) = \nabla_x \tfrac{1}{2}\|x - Z_n\|_H^2 = H(x - Z_n) = (x - Z_{n+1}).$$

Because then we get

$$X_{n+1} = X_n - \alpha_n \nabla f_{n+1}(X_n) = X_n - \tfrac{1}{n+1}(X_n - Z_{n+1}).$$

By the previous exercise, this is therefore the optimal step size. □

(iii) Prove for this quadratic loss, that SGD can be written as GD plus accumulated error

$$X_n - x_* = (x_n - x_*) - \sum_{k=0}^{n-1} \alpha_k \left(\prod_{i=k+1}^{n-1} (1 - \alpha_i H) \right) \epsilon_{k+1}. \qquad \text{(2 pts)}$$

(iv) Consider the previous setting with constant step sizes $\alpha_n = \alpha$. Additionally we are going to assume f is a quadratic loss with $H = \mathbb{I}$. Prove

$$X_n = (1-\alpha)^n x_0 + \sum_{k=1}^{n} \alpha(1-\alpha)^{n-k} Z_k.$$

Compare the estimate X_n to the mean $\bar{Z}_n$. (2 pts)