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 0 (Voluntary exercises) : Problems and Solutions

Reminder. For a real random variable X we call (if existent)

$$\varphi_X(\lambda) \coloneqq \log \mathbb{E}[e^{\lambda X}] \quad \text{and } \widetilde{\varphi}_X(\lambda) \coloneqq \log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]$$

the *logarithmic moment generating function* or shortly *log-moment generating function* and the *centered logarithmic moment generating function* or shortly *centered log-moment generating function*, respectively. Note that $\varphi_X, \widetilde{\varphi}_X \colon \mathbb{R} \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ and we denote their *domains* by

$$\mathrm{dom}(\varphi_X) \coloneqq \{\lambda \in \mathbb{R} : \varphi_X(\lambda) < \infty\} = \mathrm{dom}(\widetilde{\varphi}_X) \coloneqq \{\lambda \in \mathbb{R} : \widetilde{\varphi}_X(\lambda) < \infty\}.$$

We call

$$\varphi_X^\star(t) \coloneqq \sup_{\lambda > 0}\{\lambda t - \varphi_X(\lambda)\} \quad \text{and } \widetilde{\varphi}_X^\star(t) \coloneqq \sup_{\lambda > 0}\{\lambda t - \widetilde{\varphi}_X(\lambda)\}$$

the *Cramer transform* of φ_X and $\widetilde{\varphi}_X$, respectively. We call

$$\varphi_X^*(t) \coloneqq \sup_{\lambda \in \mathbb{R}}\{\lambda t - \varphi_X(\lambda)\} \quad \text{and } \widetilde{\varphi}_X^*(t) \coloneqq \sup_{\lambda \in \mathbb{R}}\{\lambda t - \widetilde{\varphi}_X(\lambda)\}$$

the *Legendre transform* (or *Legendre-Fenchel transform* or *convex conjugate*) of φ_X and $\widetilde{\varphi}_X$, respectively. We call a real random variable *sub-Gaussian* with parameter σ^2 (or shortly σ^2-sub-Gaussian) if $\widetilde{\varphi}_X(\lambda) \leq \frac{\sigma^2\lambda^2}{2}$ for all $\lambda \in \mathbb{R}$.

Finally, recall *Hölder's inequality.* For this consider $p, q \in [1, \infty]$ such that $\frac{1}{p} + \frac{1}{q} = 1$, where $\frac{1}{\infty} \coloneqq 0$. For two real random variables X and Y it holds that

$$\mathbb{E}[|XY|] \leq \mathbb{E}[|X|^p]^{\frac{1}{p}} \cdot \mathbb{E}[|Y|^q]^{\frac{1}{q}}.$$

Q1. (Logarithmic moment generating function of a Gaussian) Compute the centered and non centered log-moment generating function and their Legendre transforms of a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ with mean μ and variance σ^2.
Solution: Will be added once the Assignement 1 is handed in.

Q2. (Stability of sub-Gaussianity) Let X_1 and X_2 be sub-Gaussian with parameters σ_1^2 and σ_2^2, respectively.

(a) Show that αX_1 is sub-Gaussian with parameter $\alpha^2\sigma_1^2$ for a constant $\alpha \in \mathbb{R}$.
Solution: Note that $\widetilde{\varphi}_{\alpha X}(\lambda) = \log \mathbb{E}[e^{\lambda(\alpha X - \mathbb{E}[\alpha X])}] = \widetilde{\varphi}_X(\alpha\lambda) \leq \frac{\sigma^2\alpha^2\lambda^2}{2}$.

(b) If X_1 and X_2 are independent, show that X_1+X_2 is sub-Gaussian with parameter $\sigma_1^2+\sigma_2^2$.
Solution: The independence implies

$$\mathbb{E}[e^{\lambda(X_1+X_2-\mathbb{E}[X_1+X_2])}] = \mathbb{E}[e^{\lambda(X_1-\mathbb{E}X_1)}]\mathbb{E}[e^{\lambda(X_2-\mathbb{E}X_2)}].$$

Hence, $\widetilde{\varphi}_{X_1+X_2}(\lambda) = \widetilde{\varphi}_{X_1}(\lambda) + \widetilde{\varphi}_{X_2}(\lambda) \leq \frac{\sigma_1^2\lambda^2}{2} + \frac{\sigma_2^2\lambda^2}{2} = \frac{(\sigma_1^2+\sigma_2^2)\lambda^2}{2}$.

(c) Show that in general (without assuming independence), the random variable $X_1 + X_2$ is sub-Gaussian with parameter $2(\sigma_1^2 + \sigma_2^2)$. Next, show that $X_1 + X_2$ is sub-Gaussian with parameter $(\sigma_1 + \sigma_2)^2$, which improves the result.

Hint: Use Cauchy-Schwarz and Hoelder's inequality respectively.

Solution: By Cauchy-Schwarz we have

$$\mathbb{E}[e^{\lambda(X_1+X_2-\mathbb{E}[X_1+X_2])}] \le \mathbb{E}\left[e^{2\lambda(X_1-\mathbb{E}X_1)}\right]^{\frac{1}{2}} \mathbb{E}\left[e^{2\lambda(X_2-\mathbb{E}X_2)}\right]^{\frac{1}{2}}.$$

Hence, we can estimate

$$\widetilde{\varphi}_{X_1+X_2}(\lambda) \le \frac{\widetilde{\varphi}_{X_1}(2\lambda) + \widetilde{\varphi}_{X_2}(2\lambda)}{2} \le \frac{4\sigma_1^2\lambda^2 + 4\sigma_2^2\lambda^2}{4} = \frac{2(\sigma_1^2+\sigma_2^2)\lambda^2}{2}.$$

By Jensen's inequality with $p = \frac{\sigma_1+\sigma_2}{\sigma_1}$ and $q = \frac{\sigma_1+\sigma_2}{\sigma_2}$ we have

$$\mathbb{E}[e^{\lambda(X_1+X_2-\mathbb{E}[X_1+X_2])}] \le \mathbb{E}\left[e^{\frac{\sigma_1+\sigma_2}{\sigma_1}\lambda(X_1-\mathbb{E}X_1)}\right]^{\frac{\sigma_1}{\sigma_1+\sigma_2}} \cdot \mathbb{E}\left[e^{\frac{\sigma_1+\sigma_2}{\sigma_2}\lambda(X_2-\mathbb{E}X_2)}\right]^{\frac{\sigma_2}{\sigma_1+\sigma_2}}.$$

Hence, we can estimate

$$\begin{aligned}
\widetilde{\varphi}_{X_1+X_2}(\lambda) &\le \frac{\sigma_1}{\sigma_1+\sigma_2}\cdot\widetilde{\varphi}_{X_1}\left(\frac{\sigma_1+\sigma_2}{\sigma_1}\lambda\right) + \frac{\sigma_2}{\sigma_1+\sigma_2}\cdot\widetilde{\varphi}_{X_2}\left(\frac{\sigma_2+\sigma_2}{\sigma_2}\lambda\right) \\
&\le \frac{\sigma_1}{\sigma_1+\sigma_2}\cdot\frac{(\sigma_1+\sigma_2)^2\lambda^2\sigma_1^2}{2\sigma_1^2} + \frac{\sigma_2}{\sigma_1+\sigma_2}\cdot\frac{(\sigma_1+\sigma_2)^2\lambda^2\sigma_2^2}{2\sigma_2^2} \\
&= \frac{(\sigma_1(\sigma_1+\sigma_2) + \sigma_2(\sigma_1+\sigma_2))\cdot\lambda^2}{2} \\
&= \frac{(\sigma_1+\sigma_2)^2\lambda^2}{2}.
\end{aligned}$$

Recall Young's inequality $2\sigma_1\sigma_2 \le \sigma_1^2 + \sigma_2^2$ which yields $(\sigma_1+\sigma_2)^2 \le 2(\sigma_1^2+\sigma_2^2)$.

Q3. (Properties of logarithmic moment generating functions) For this exercise you can assume $\mathbb{E}X = 0$ in which case $\varphi_X = \widetilde{\varphi}_X$, but the statements hold in general.

(a) *Convexity.* Show that φ_X and $\widetilde{\varphi}_X$ are convex functions.

Hint: Hölder's inequality

Solution: Consider $\lambda_1, \lambda_2 \in \mathbb{R}$ and $t \in (0,1)$. Then by Hölder's inequality we have $\mathbb{E}[YZ] \le \mathbb{E}[Y^{\frac{1}{t}}]^t \mathbb{E}[Z^{\frac{1}{1-t}}]^{1-t}$ for $Y, Z \ge 0$ and hence we find

$$\begin{aligned}
\varphi_X(t\lambda_1 + (1-t)\lambda_2) &= \log\left(\mathbb{E}[e^{(t\lambda_1+(1-t)\lambda_2)X}]\right) = \log\left(\mathbb{E}[e^{t\lambda_1 X}e^{(1-t)\lambda_2 X}]\right) \\
&\le \log\left(\mathbb{E}[e^{\lambda_1 X}]^t\mathbb{E}[e^{\lambda_2 X}]^{1-t}\right) = t\varphi_X(\lambda_1) + (1-t)\varphi_X(\lambda_2).
\end{aligned}$$

(b) *Semi-continuity.* Show that φ_X and $\widetilde{\varphi}_X$ lower semi-continuous, i.e., if $\lambda_n \to \lambda$ for $n \to \infty$ then

$$\liminf_{n\to\infty} \varphi_X(\lambda_n) \ge \varphi_X(\lambda) \quad \text{and} \quad \liminf_{n\to\infty} \widetilde{\varphi}_X(\lambda_n) \ge \widetilde{\varphi}_X(\lambda).$$

Hint: Fatou's lemma

Solution: Using Fatou's lemma we find

$$\liminf_{n\to\infty} \varphi_X(\lambda_n) = \log\left(\liminf_{n\to\infty} \mathbb{E}[e^{\lambda_n X}]\right) \ge \log\left(\mathbb{E}[\liminf_{n\to\infty} e^{\lambda_n X}]\right) = \log\mathbb{E}[e^{\lambda X}] = \varphi_X(\lambda).$$

(c) *Existence of moments.* Assume that $\varphi_X(\lambda) < \infty$ or $\widetilde{\varphi}_X(\lambda) < \infty$ for all $\lambda \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. Show that all moments exist, i.e., $\mathbb{E}[|X|^k] < \infty$ for all $k \in \mathbb{N}$.
Solution: Note that $e^{\lambda|X|} \le e^{\lambda X} + e^{-\lambda X}$ and hence for $\lambda \in (0, \varepsilon)$ we have

$$\mathbb{E}[e^{\lambda|X|}] \le \varphi_X(\lambda) + \varphi_X(-\lambda) < \infty.$$

In particular, this implies that

$$\frac{\lambda^k \mathbb{E}[|X|^k]}{k!} \le \mathbb{E}[e^{\lambda|X|}] < \infty.$$

(d) *Smoothness.* Show that φ_X and $\widetilde{\varphi}_X$ are smooth, i.e., infinitely many times continuously differentiable, functions on the interior of their domains.
Hint: It suffices to show that e^{φ_X} and $e^{\widetilde{\varphi}_X}$ are smooth for which you can use the dominated convergence theorem.
Solution: It suffices to show smoothness of $M(\lambda) \coloneqq e^{\varphi_X(\lambda)} = \mathbb{E}[e^{\lambda X}]$ for which we fix $\lambda \in \operatorname{int}(\operatorname{dom}(\varphi_X))$. Recall that integration and differentiation can be exchanged if the is a dominating integrable function (locally) independent of the variable that we consider in the differentiation. In our case this means that we need to find an upper bound $|\partial_\lambda^{(k)} e^{\lambda' X}| = |X|^k e^{\lambda' X}$ with finite expectation that holds for all $\lambda' \in (\lambda - \varepsilon, \lambda + \varepsilon)$. First, note that

$$|X|^k e^{\lambda' X} \le |X|^k \left(e^{(\lambda+\varepsilon)X} + e^{(\lambda-\varepsilon)X} \right)$$

and hence it remains to show that $\mathbb{E}[|X|^k e^{(\lambda \pm \varepsilon)X}] < \infty$. Note that $(\lambda \pm \varepsilon)p \in \operatorname{dom}(\varphi_X)$ for $p = (1 - 1/l)^{-1}$ for $l \in \mathbb{N}$ large enough. Now, we can use Hölder's inequality to show

$$\mathbb{E}[|X|^k e^{(\lambda \pm \varepsilon)X}] \le \mathbb{E}[|X|^{kl}]^{\frac{1}{l}} \cdot \mathbb{E}[e^{p(\lambda \pm \varepsilon)X}]^{\frac{1}{p}} < \infty$$

since all absolut moments of X exist.

(e) *Derivatives.* Show that

$$\varphi_X'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{M(\lambda)} \quad \text{and } \widetilde{\varphi}_X'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda(X - \mathbb{E}X)}]}{\widetilde{M}(\lambda)},$$

where $M(\lambda) \coloneqq \mathbb{E}[e^{\lambda X}] = e^{\varphi_X(\lambda)}$ and $\widetilde{M}(\lambda) \coloneqq \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] = e^{\widetilde{\varphi}_X(\lambda)}$ denote the expontial and centered moment generating functions of X.
Solution: We have seen in the previous exercise that we can exchange integration and differentiation by the dominated convergence theorem. Hence, we can compute

$$\varphi_X'(\lambda) = \partial_\lambda \log \mathbb{E}[e^{\lambda X}] = \frac{\mathbb{E}[\partial_\lambda e^{\lambda X}]}{E[e^{\lambda X}]} = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}$$

and analogously for $\widetilde{\varphi}_X$.

(f) *Cramer transform equals Legendre transform.* Assume that $\widetilde{\varphi}_X$ is finite on $(-\varepsilon, \varepsilon)$. Show that

$$\widetilde{\varphi}_X^\star(t) = \sup_{\lambda > 0} \{\lambda t - \widetilde{\varphi}_X(\lambda)\} = \sup_{\lambda \in \mathbb{R}} \{\lambda t - \widetilde{\varphi}_X(\lambda)\} = \widetilde{\varphi}_X^*(t) \quad \text{for all } t \ge 0.$$

Solution: It is clear from the definition that $\widetilde{\varphi}_X^\star \le \widetilde{\varphi}_X^*$. The function $\widetilde{\varphi}_X$ is smooth and in particular continuous and hence

$$\widetilde{\varphi}^\star(t) = \sup_{\lambda > 0} \{\lambda t - \widetilde{\varphi}_X(\lambda)\} = \sup_{\lambda \ge 0} \{\lambda t - \widetilde{\varphi}_X(\lambda)\}.$$

Note that $\widetilde{\varphi}_X(0) = 0$ and therefore $\widetilde{\varphi}_X^\star(0) \geq 0 \cdot t - \widetilde{\varphi}_X(0) = 0$. By Jensen's inequality we have $\widetilde{\varphi}_X(\lambda) \geq \mathbb{E}[\lambda X] = 0$ for all $\lambda \in \mathbb{R}$ and hence for $\lambda < 0$ and $t \geq 0$ we have

$$\lambda t - \widetilde{\varphi}_X(\lambda) \leq 0,$$

which shows the desired equality.