This homework is due at **2pm, September 30, 2019**. There are in total 105 points. Your actual grade of this homework will be $\min\{100, \text{points you get}\}$. If you have any problem about this solution, please contact **weicheng.frank.lee@gmail.com**.

Problem 1

Let $f : \mathbb{R}^p \to \mathbb{R}$. Its *gradient* is a p-dimensional vector given by

$$\nabla f(x) := \left(\frac{\partial f}{\partial x^{(1)}}(x), \ldots, \frac{\partial f}{\partial x^{(p)}}(x) \right), \quad \forall x \in \mathbb{R}^p,$$

where $x^{(i)}$ denotes the i-th entry of the vector x. Its *Hessian* is a matrix in $\mathbb{R}^{p \times p}$ given by

$$\left[\nabla^2 f(x) \right]^{(i,j)} := \frac{\partial^2 f}{\partial x^{(i)} \partial x^{(j)}}(x), \quad \forall x \in \mathbb{R}^p,$$

for all $1 \le i, j \le p$, where $\left[\nabla^2 f(x) \right]^{(i,j)}$ denotes the (i, j)-th entry of the matrix $\nabla^2 f(x)$.

Let $a \in \mathbb{R}^p$. A machine learning algorithm called *logistic regression* requires minimizing a sum of functions of the form

$$g(x) := \log\left(1 + \mathrm{e}^{-\langle a, x \rangle}\right), \quad \forall x \in \mathbb{R}^p.$$

1. (15 points) **Show that**

$$\nabla g(x) = \frac{-a}{1 + \mathrm{e}^{\langle a, x \rangle}}, \quad \forall x \in \mathbb{R}^p.$$

Solution. $\forall i \in [p]$,

$$\nabla g(x) e_i = \frac{\partial g(x)}{\partial x^{(i)}} = \frac{\mathrm{e}^{-\langle a, x \rangle} \left(-\langle a, x \rangle\right)'}{1 + \mathrm{e}^{-\langle a, x \rangle}} = \frac{-a_i \mathrm{e}^{-\langle a, x \rangle}}{1 + \mathrm{e}^{-\langle a, x \rangle}} = \frac{-a_i}{1 + \mathrm{e}^{\langle a, x \rangle}}.$$

where e_i denotes the standard basis of i-th dimension.

2. (15 points) **Show that**

$$\nabla^2 g(x) = \frac{\mathrm{e}^{\langle a, x \rangle} a a^{\mathrm{T}}}{\left(1 + \mathrm{e}^{\langle a, x \rangle}\right)^2}, \quad \forall x \in \mathbb{R}^p,$$

where a^{T} denotes the transpose of a.

Solution. $\forall \left(i, j\right) \in [p] \times [p]$,

$$e_i^{\mathrm{T}} \nabla^2 g(x) e_j = \frac{\partial^2 g(x)}{\partial x^{(i)} \partial x^{(j)}} = \frac{0 \times \left(1 + \mathrm{e}^{\langle a, x \rangle}\right) + a_j a_i \mathrm{e}^{\langle a, x \rangle}}{\left(1 + \mathrm{e}^{\langle a, x \rangle}\right)^2} = e_i^{\mathrm{T}} \frac{\mathrm{e}^{\langle a, x \rangle} a a^{\mathrm{T}}}{\left(1 + \mathrm{e}^{\langle a, x \rangle}\right)^2} e_j.$$

3. (15 points) Let $A, B \in \mathbb{R}^{p \times p}$. We write $A \ge B$ if and only if $(A - B)$ is positive semi-definite, and $A \le B$ if and only if $B \ge A$. **Show that**

$$0 \le \nabla^2 g(x) \le \frac{\|a\|_2^2}{4} I, \quad \forall x \in \mathbb{R}^p,$$

where I denotes the identity matrix.

Solution. Observe that the inequalities are equivalent to

$$\lambda_{min}(\nabla^2 g(x)) \ge 0, \quad \forall x \in \mathbb{R}^p.$$

$$\lambda_{max}(\nabla^2 g(x)) \le \frac{\|a\|_2^2}{4}, \quad \forall x \in \mathbb{R}^p.$$

Since $\nabla^2 g(x)$ has only two eigenvalues $\{0, \frac{\mathrm{e}^{\langle a, x \rangle}}{(1+\mathrm{e}^{\langle a, x \rangle})^2} \|a\|_2^2\}$ and $\mathrm{e}^{\langle a, x \rangle} > 0 \quad \forall x \in \mathbb{R}^p$, then the first inequality is trivial and the second is followed by

$$\frac{t}{(1+t)^2} \le \frac{1}{4}, \quad \forall t > 0.$$

Problem 2

Let ξ be a random variable taking values in $\{-1,1\}$. Define

$$\varphi(\beta) := \log\left(\mathsf{E}\,e^{\beta\xi}\right), \quad \forall \beta \in \mathbb{R},$$

where $\mathsf{E}\,e^{\beta\xi}$ denotes the expectation of $e^{\beta\xi}$.
Useful Fact.

1. Since $\mathsf{E}\,e^{\beta\xi} < \infty \quad \forall \beta \in \mathbb{R}$, we have $\frac{d^n \mathsf{E}\,e^{\beta\xi}}{d\beta^n} = \mathsf{E}\,\xi^n e^{\beta\xi} \quad \forall \beta \in \mathbb{R}, \forall n \in \mathbb{N}^+$.
2. $\varphi'(\beta) = \frac{\mathsf{E}\,\xi e^{\beta\xi}}{\mathsf{E}\,e^{\beta\xi}} \quad \forall \beta \in \mathbb{R}$, $\varphi'(0) = \mathsf{E}\,\xi$ and $\varphi(0) = 0$.
3. $|\varphi'(\beta)| \le 1 \forall \beta \in \mathbb{R}$. In particular, $\varphi'(\beta) = 1 \Leftrightarrow P(\{\xi = 1\}) = 1$ and $\varphi'(\beta) = -1 \Leftrightarrow P(\{\xi = -1\}) = 1$.
4. Let η be a random variable taking values in $\{-1,1\}$. If $P\left(\{\eta = 1\}\right) = p$, then $\mathsf{E}\,\eta = 2p-1, \mathsf{E}\,\eta^2 = 1$ and $\operatorname{var}\eta = -4p^2 + 4p$.

1. (15 points) **Show that**

$$\varphi''(\beta) = \mathsf{E}\left[\left(\eta_\beta - \mathsf{E}\,\eta_\beta\right)^2\right], \quad \varphi'''(\beta) = \mathsf{E}\left[\left(\eta_\beta - \mathsf{E}\,\eta_\beta\right)^3\right], \quad \forall \beta \in \mathbb{R},$$

for some random variable η_β taking values in $\{-1,1\}$ whose probability distribution depends on β.

Solution.

$$\varphi''(\beta) = \frac{\mathsf{E}\,\xi^2 e^{\beta\xi}\,\mathsf{E}\,e^{\beta\xi} - \left(\mathsf{E}\,\xi e^{\beta\xi}\right)^2}{\left(\mathsf{E}\,e^{\beta\xi}\right)^2} = 1 - \left(\frac{\mathsf{E}\,\xi e^{\beta\xi}}{\mathsf{E}\,e^{\beta\xi}}\right)^2 = 1 - \varphi'(\beta)^2.$$

$$\varphi'''(\beta) = -2\varphi'(\beta)\varphi''(\beta).$$

Let a=$\varphi'(\beta)$, By fact 4, we can get η_β by solving

$$\varphi''(\beta) = 1 - a^2 = -4p^2 + 4p = \operatorname{var}\eta_\beta.$$

By fact 3, η_β is well-defined, $P\left(\{\eta_\beta = 1\}\right) = \frac{1+a}{2}$ and $\mathsf{E}\,\eta_\beta = \varphi'(\beta)$. It remains to check $\varphi'''(\beta) = \mathsf{E}\left[\left(\eta_\beta - \mathsf{E}\,\eta_\beta\right)^3\right]$.

$$\mathsf{E}\left[\left(\eta_\beta - \mathsf{E}\,\eta_\beta\right)^3\right] = \mathsf{E}\left[\eta_\beta^3 - 3\eta_\beta^2 \mathsf{E}\,\eta_\beta + 3\eta_\beta \mathsf{E}\left[\eta_\beta\right]^2 - \mathsf{E}\left[\eta_\beta\right]^3\right] = -2\mathsf{E}\,\eta_\beta\left(1 - \mathsf{E}\left[\eta_\beta\right]^2\right) = -2\varphi'(\beta)\varphi''(\beta) = \varphi'''(\beta).$$

2. (15 points) **Show that**

$$\varphi''(\gamma) \le e^{2|\gamma-\beta|}\varphi''(\beta), \quad \forall \beta, \gamma \in \mathbb{R}.$$

HINT: By the results above, we have

$$\varphi'''(\beta) \le 2\varphi''(\beta), \quad \forall \beta \in \mathbb{R}.$$

Solution. By fact 3, we can assume $\varphi''(\beta) \ne 0$ and $\varphi''(\gamma) \ne 0$. Let $\phi(\beta) = \log\left(\varphi''(\beta)\right)$, we have $\phi''(\beta) = \frac{\varphi'''(\beta)}{\varphi''(\beta)}$.
By Mean Value Theorem, for a point c in the line segment created by $\{\gamma, \beta\}$, we have

$$|\phi''(\gamma) - \phi''(\beta)| = |\gamma - \beta||\phi'(c)| \le 2|\gamma - \beta|.$$

That is to say, $\log\left(\frac{\varphi''(\gamma)}{\varphi''(\beta)}\right) \le 2|\gamma - \beta|$, so

$$\varphi''(\gamma) \le e^{2|\gamma-\beta|}\varphi''(\beta), \quad \forall \beta, \gamma \in \mathbb{R}.$$

Problem 2 continued on the next page...

3. (15 points) **Show that**

$$\varphi'(\gamma) \le \varphi'(\beta) + \left[\frac{e^{2(\gamma-\beta)}-1}{2\left(\gamma-\beta\right)}\right]\varphi''(\beta)(\gamma-\beta), \quad \forall \beta, \gamma \in \mathbb{R} \textbf{ such that } \gamma > \beta.$$

Solution.

$$\begin{aligned}
\varphi''(\gamma) &= \varphi''(\beta) + \int_\beta^\gamma \varphi''(s)ds \\
&\overset{(2)}{\le} \varphi''(\beta) + \int_\beta^\gamma e^{2(s-\beta)}\varphi''(\beta)ds \\
&= \varphi'(\beta) + \left[\frac{e^{2(\gamma-\beta)}-1}{2\left(\gamma-\beta\right)}\right]\varphi''(\beta)(\gamma-\beta).
\end{aligned}$$

4. (15 points) **Use the results above to prove that**

$$\log\left[\mathsf{E}\, e^{\lambda(\xi-\mathsf{E}\,\xi)}\right] \le \frac{h(2\lambda)}{4}\mathsf{var}\,\xi, \quad \forall \lambda > 0,$$

where $h(x) := e^x - x - 1$ **and** $\mathsf{var}\,\xi$ **denotes the variance of** ξ**.** This is essentially *Bennett's inequality*. See, e.g., [1, Theorem 2.9] for the details; however, notice we want a proof based on the results above and do not copy the proof in [1].

HINT: Compare $\varphi(\lambda)$ and $\varphi(0)$.

Solution. Observe that $\forall \lambda > 0$

$$\log\left[\mathsf{E}\, e^{\lambda(\xi-\mathsf{E}\,\xi)}\right] = \log\left[\mathsf{E}\, e^{\lambda(\xi)}/\mathsf{E}\, e^{\lambda(\mathsf{E}\,\xi)}\right] = \log(\mathsf{E}\, e^{\lambda(\xi)}) - \lambda\mathsf{E}\,\xi = \varphi(\lambda) - \lambda\varphi'(0).$$

The inequality follows by Taylor Expansion and $\varphi''(0) = 1 - \varphi'(0)^2 = \mathsf{var}\,\xi$,

$$\begin{aligned}
\varphi(\lambda) &= \varphi(0) + \varphi'(0)\lambda + \int_0^\lambda \varphi''(s)(\lambda-s)ds \\
&\overset{(2)}{\le} \varphi(0) + \varphi'(0)\lambda + \int_0^\lambda e^{2s}\varphi''(0)(\lambda-s)ds \\
&= \varphi(0) + \varphi'(0)\lambda + \varphi''(0)\left(\frac{e^{2\lambda-2\lambda-1}}{4}\right) \\
&= \varphi'(0)\lambda + \frac{h(2\lambda)}{4}\mathsf{var}\,\xi.
\end{aligned}$$

References

[1] Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford Univ. Press, Oxford, 2013.