Answer to 2.1:

(a) We show the result for \( x > x_0 \) for simplicity. Following the hint, we write

\[
x = \frac{x - x_0}{y - x_0} y + \frac{y - x}{y - x_0} x_0,
\]

which gives

\[
f(x) \leq \frac{x - x_0}{y - x_0} f(y) + \frac{y - x}{y - x_0} f(x_0) = \frac{x - x_0}{y - x_0} f(y) + f(x_0) - \frac{x - x_0}{y - x_0} f(x_0).
\]

Rearranging, we obtain

\[
f(x) - f(x_0) \leq \frac{x - x_0}{y - x_0} (f(y) - f(x_0)),
\]

which is equivalent to \( s(x) \leq s(y) \) because \( x > x_0 \). The case when \( y \leq x < x_0 \) is completely similar.

(b) First, let \( 0 < t_1 \leq t_2 \). Then

\[
\frac{f(x + t_1v) - f(x)}{t_1} = \frac{f(x + t_2(t_1/t_2)v) - f(x)}{t_1} = \frac{f((1 - t_1/t_2)x + (t_1/t_2)(x + t_2v)) - f(x)}{t_1} \leq \frac{(1 - t_1/t_2)f(x) + (t_1/t_2)f(x + t_2v) - f(x)}{t_1} = \frac{f(x + t_2v) - f(x)}{t_2}.
\]

That is, \( t \mapsto \frac{f(x + tv) - f(x)}{t} \) is non-decreasing, and so the limit as \( t \downarrow 0 \) of the quantity exists and is the infimum.

(c) Because of part (a), we know that the slope function is increasing, and thus \( f(x - t) - f(x) \leq f(x + t) - f(x) \), so that \( f'(x; -1) \leq f'(x; 1) \).

(d) We have that \( f(x + t) \geq f(x) + tf'(x; 1) \) for all \( t \geq 0 \) by definition of the directional derivative, and similarly, \( f(x - t) \geq f(x) + tf'(x; -1) \) by an identical calculation. If \( y \geq x \), we obtain

\[
f(y) \geq f(x) + (y - x)f'(x; -1) \geq f(x) + (y - x)f'(x; -1)
\]

by the first inequality, because \( f'(x; 1) \geq f'(x; -1) \) and \( y - x \). If \( y \leq x \), we set \( t = x - y \) in the second quantity to obtain

\[
f(y) \geq f(x) + (y - x)f'(x; -1) \geq f(x) + (y - x)f'(x; 1)
\]

because \( y - x \leq 0 \). As \( g \in [f'(x; -1), f'(x; 1)] \), we have the result.

(e) First, we argue that if \( f \) is strictly convex at the point \( x \), then for any \( g \in \partial f(x) \), either

i. \( f(y) > f(x) + g(y - x) \) for all \( y > x \)

ii. \( f(y) > f(x) + g(y - x) \) for all \( y < x \)
Indeed, assume neither of these is the case, that is, we have for some \( y_0 < x < y_1 \) that 
\[ f(y_i) = f(x) + g(y_i - x) \quad \text{for } i = 0, 1. \]
Define the slope function
\[ s(y) := \frac{f(y) - f(x)}{y - x}. \]
Then we must have \( s(y_1) = f'(x; 1) = g \) by the criterion of increasing slopes and part (b), and 
\[ s(y_0) = -f'(x; -1) = g \]
by the same argument, so \( f'(x; 1) = -f'(x; -1) = g \). Moreover, we have 
\[ s(y) = f'(x; 1) \]
for all \( y \in (x, y_1] \) and \( s(y) = -f'(x; -1) \) for all \( y \in [y_0, x) \). In particular, we have 
that \( f \) is linear on the intervals \([y_0, x]\) and \([x, y_1]\), with slope \( g \), so that 
\[ f(y) = f(x) + g(y - x) \]
for all \( y \in [y_0, y_1] \). Clearly this function is not strictly convex at \( x \), which is a contradiction.

Now we prove the assertion of the question. Let \( \mathbb{E}[X] \in \text{int dom } f \). Taking 
\( g \in \partial f(x) \) and assuming \( \mathbb{P}(X = \mathbb{E}[X]) < 1 \), there must be some \( \epsilon > 0 \)
such that \( \mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) > 0 \), and so
\[ f(X) > f(\mathbb{E}[X]) + g(X - \mathbb{E}[X]) \]
with positive probability. Thus \( \mathbb{E}[f(X)] > f(\mathbb{E}[X]) + g\mathbb{E}[X - \mathbb{E}[X]] = f(\mathbb{E}[X]) \). Conversely, 
if \( X = \mathbb{E}[X] \) with probability 1, then clearly \( f(\mathbb{E}[X]) = \mathbb{E}[f(X)] \). Lastly, if \( \mathbb{E}[X] \notin \text{int dom } f \), 
then if \( X \notin \mathbb{E}[X] \) with positive probability, we have \( X \notin \text{dom } f \) with positive probability and 
\( \mathbb{E}[f(X)] = \infty \).

Answer to 2.2: We use Jensen’s inequality is as follows: if \( h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\} \) is a convex 
function and \( X \) a random vector, then \( \mathbb{E}[h(X)] \geq h(\mathbb{E}[X]) \). If \( h \) is strictly convex at \( \mathbb{E}[X] \), then the 
inequality is strict unless \( X = \mathbb{E}[X] \) with probability 1.

We have that \( t \mapsto -\log t \) is a strictly convex function, so that for any random variable \( Y \geq 0 \) we 
have \( \mathbb{E}[-\log Y] \geq -\log \mathbb{E}[Y] \) with equality if and only if \( Y = \mathbb{E}[Y] \) with probability 1. Define the 
random variable \( L(X) = \frac{q(X)}{p(X)} \), which exists and is finite \( P \)-a.s. Then \( \mathbb{E}_P[L(X)] = \int q(x) d\mu(x) = 1 \), 
and \( D_1(P|Q) = \mathbb{E}_P[-\log L(X)] \geq -\log \mathbb{E}_P[L(X)] = 0 \), with strict inequality unless \( L(X) = 1 \) 
with \( P \)-probability 1, or \( P = Q \).

Answer to 2.3:
(a) Following the hint, we let \( \lambda \in [0, 1] \) and define the conjugate pair \( p = 1/\lambda \) and \( q = 1/(1 - \lambda) \).
We then write
\[
A(\lambda \theta_0 + (1 - \lambda) \theta_1) = \log \int \exp (\lambda \langle \theta_0, T(x) \rangle + (1 - \lambda) \langle \theta_1, T(x) \rangle) \, d\mu(x)
\]
\[
= \log \int (e^{\langle \theta_0, T(x) \rangle})^{\frac{1}{p}} (e^{\langle \theta_1, T(x) \rangle})^{\frac{1}{q}} d\mu(x)
\]
\[
\leq \log \left( \int e^{\langle \theta_0, T(x) \rangle} d\mu(x) \right)^{\frac{1}{p}} \left( \int e^{\langle \theta_1, T(x) \rangle} d\mu(x) \right)^{\frac{1}{q}}
\]
by Hölder’s inequality. Noting that \( \log ab = \log a + \log b \) we obtain
\[
A(\lambda \theta_0 + (1 - \lambda) \theta_1) \leq \frac{1}{p} \log \int e^{\langle \theta_0, T(x) \rangle} d\mu(x) + \frac{1}{q} \log \int e^{\langle \theta_1, T(x) \rangle} d\mu(x),
\]
which is our desired result.
(b) We compute the KL-divergence of $P_{\theta_0}$ and $P_{\theta_1}$, which is

$$D_{\text{KL}} (P_{\theta_0} \| P_{\theta_1}) = A(\theta_1) - A(\theta_0) - \langle \nabla A(\theta_0), \theta_1 - \theta_0 \rangle.$$  

We know that for some $\tilde{\theta} \in [\theta_0, \theta_1]$, we have

$$A(\theta_1) = A(\theta_0) + \langle \nabla A(\theta_0), \theta_1 - \theta_0 \rangle + \frac{1}{2} \frac{(\theta_1 - \theta_0)^T \nabla^2 A(\tilde{\theta})(\theta_1 - \theta_0)}{>0} ,$$

by a Taylor expansion, and the strict positivity of $\nabla^2 A(\theta)$ gives that $D_{\text{KL}} (P_{\theta_0} \| P_{\theta_1}) > 0$.

\[\square\]

**Answer to 2.5:** We answer each part in turn.

(a) This is trivial: for any measure $\mu$ on $X$ we have by convexity of $\ell$ that for any $t \in [0,1]$,

$$\int \ell(t\theta+(1-t)\theta',x) d\mu(x) \leq \int t\ell(\theta,x)+(1-t)\ell(\theta',x) d\mu(x) = t \int \ell(\theta,x) d\mu(x) + (1-t)\ell(\theta',x) d\mu(x)$$

for any $\theta, \theta'$. Take $\mu$ to be either $P$ or the empirical distribution on $X_1, \ldots, X_n$ to get the result.

(b) Let $\hat{P}_n$ denote the empirical distribution on $X_1, \ldots, X_n$, so that $E_{\hat{P}_n} [f(X)] = \frac{1}{n} \sum_{i=1}^n f(X_i)$. We know that $\theta^*$ satisfies $\nabla R(\theta^*) = 0$, so that

$$E[\nabla \ell(\theta^*, X)] = 0.$$ 

Thus, we have that

$$\sqrt{n}E_{\hat{P}_n} [\nabla \ell(\theta^*, X)] \overset{d}{\to} N(0, \Sigma) \text{ and } E_{\hat{P}_n} [\nabla \ell(\theta^*, X)] \overset{a}{\to} 0$$

by the multivariate central limit theorem. Moreover, we have $E[\nabla^2 \ell(\theta^*, X)] = \nabla^2 R(\theta^*) > 0$, and the SLLN implies that $E_{\hat{P}_n} [\nabla^2 \ell(\theta^*, X)] \overset{a}{\to} \nabla^2 R(\theta^*)$. Let $\lambda > 0$ be the minimal eigenvalue of $\nabla^2 R(\theta^*)$ so that $\nabla^2 R(\theta^*) \succeq \lambda I$. Then with probability 1, for all sufficiently large $n$ we have $E_{\hat{P}_n} [\nabla^2 \ell(\theta^*, X)] \succeq \frac{\lambda}{4} I$. Let $H = E[H^2(X)] \frac{1}{2}$ be the expectation of $H$, and assume that $N$ is large enough that $\frac{1}{n} \sum_{i=1}^n H^2(X_i) \leq 4H^2$ for all $n \geq N$ (again, this is possible by the SLLN). Then we have

$$\nabla^2 \hat{R}_n(\theta) \succeq \nabla^2 \hat{R}_n(\theta^*) - \|\nabla^2 \hat{R}_n(\theta^*) - \nabla^2 \hat{R}_n(\theta)\|_{\text{op}} I$$

$$\succeq \left(\frac{3\lambda}{4} - \|\theta - \theta^*\| \left(\frac{1}{n} \sum_{i=1}^n H(X_i)\right)\right) I$$

$$\succeq \left(\frac{3\lambda}{4} - 2H \|\theta - \theta^*\|\right) I,$$

so that for all $\theta$ such that $\|\theta - \theta^*\| \leq \frac{\lambda}{8H}$ we have $\nabla^2 \hat{R}_n(\theta) \succeq \frac{\lambda}{2} I$.

Now, applying the hint on convexity, we have

$$\hat{R}_n(\theta) \geq \hat{R}_n(\theta^*) + \nabla \hat{R}_n(\theta^*)^T (\theta - \theta^*) + \frac{\lambda}{4} \min\{\|\theta - \theta^*\|^2, \frac{\lambda}{8H} \|\theta - \theta^*\|\}$$
for all \( \theta \), or, a simpler condition, for some \( c > 0 \) we have
\[
\hat{R}_n(\theta) \geq \hat{R}_n(\theta^*) + \nabla \hat{R}_n(\theta^*)^T (\theta - \theta^*) + c \min \left\{ \|\theta - \theta^*\|^2, \|\theta - \theta^*\| \right\}.
\]

Let \( \epsilon < c \) be otherwise arbitrary; we know that for sufficiently large \( n \) we also have \( \|\nabla \hat{R}_n(\theta^*)\| \leq \epsilon \). Consequently, we find by the Cauchy-Schwartz inequality that
\[
\hat{R}_n(\theta) \geq \hat{R}_n(\theta^*) - \epsilon \|\theta - \theta^*\| + c \min \{\|\theta - \theta^*\|, \|\theta - \theta^*\|^2\},
\]
and \( \|\theta - \theta^*\| > \sqrt{\epsilon/c} \) implies that
\[
- \epsilon \|\theta^* - \theta\| + c \min \{\|\theta^* - \theta\|, \|\theta^* - \theta\|^2\} > (\sqrt{\epsilon} - \epsilon) \|\theta - \theta^*\| > 0,
\]
or \( \hat{R}_n(\theta) > \hat{R}_n(\theta^*) \). Thus we must have \( \|\hat{\theta}_n - \theta^*\| \leq \sqrt{\epsilon/c} \) eventually, so \( \hat{\theta}_n \overset{a.s.}{\rightarrow} \theta^* \).

**For fun, the convexity result** For completeness, we also show the result on convexity to keep our entire argument self-contained. Let \( f \) be convex and satisfy the conditions in the theorem. Then by a Taylor expansion, for any \( \theta \) satisfying \( \|\theta - \theta_0\| \leq c \), we have for some \( \tilde{\theta} \in [\theta, \theta_0] \) that
\[
f(\theta) = f(\theta_0) + (\nabla f(\theta_0), \theta - \theta_0) + \frac{1}{2} (\theta - \theta_0)^T \nabla^2 f(\tilde{\theta})(\theta - \theta_0) \geq f(\theta_0) + (\nabla f(\theta_0), \theta - \theta_0) + \frac{\lambda}{2} \|\theta - \theta_0\|^2
\]
because \( \|\tilde{\theta} - \theta_0\| \leq c \). Now, suppose that \( \|\theta - \theta_0\| \geq c \). Let \( t \in [0,1] \) be such that \( \theta_t := t\theta + (1-t)\theta_0 \) satisfies \( \|\theta_t - \theta_0\| = c \), or \( t = \frac{c}{\|\theta - \theta_0\|} \). By convexity and the assumption of the question, we then obtain
\[
t f(\theta) + (1-t) f(\theta_0) \geq f(\theta_t) \geq f(\theta_0) + (\nabla f(\theta_0), \theta_t - \theta_0) + \frac{\lambda}{2} \|\theta_t - \theta_0\|^2.
\]
But of course, we have \( \theta_t - \theta_0 = t(\theta - \theta_0) \), and so we have
\[
t f(\theta) + (1-t) f(\theta_0) \geq f(\theta_0) + t(\nabla f(\theta_0), \theta - \theta_0) + \frac{\lambda}{2} t^2 \|\theta - \theta_0\|^2,
\]
which, rearranged and dividing by \( t = \frac{c}{\|\theta - \theta_0\|} > 0 \), yields
\[
f(\theta) \geq f(\theta_0) + (\nabla f(\theta_0), \theta - \theta_0) + \frac{\lambda}{2} t \|\theta - \theta_0\|^2 = f(\theta_0) + (\nabla f(\theta_0), \theta - \theta_0) + \frac{\lambda c}{2} \|\theta - \theta_0\|.
\]
This is our desired result, as we see that
\[
f(\theta) \geq \begin{cases} f(\theta_0) + (\nabla f(\theta_0), \theta - \theta_0) + \frac{\lambda}{2} \|\theta - \theta_0\|^2 & \text{if } \|\theta - \theta_0\| \leq c \\ f(\theta_0) + (\nabla f(\theta_0), \theta - \theta_0) + \frac{\lambda c}{2} \|\theta - \theta_0\| & \text{if } \|\theta - \theta_0\| > c. \end{cases}
\]
(c) We may assume w.l.o.g. that \( \|\theta^* - \tilde{\theta}\| \leq \epsilon \) for any \( \epsilon \), as we already have assumed that \( \hat{\theta}_n - \theta^* \overset{P}{\rightarrow} 0 \). Then by performing a Taylor expansion of \( \nabla \hat{R}_n \) around \( \theta^* \), we have
\[
0 = \nabla \hat{R}_n(\hat{\theta}_n) = \nabla \hat{R}_n(\theta^*) + \nabla^2 \hat{R}_n(\theta^*)(\hat{\theta}_n - \theta^*) + \left( \frac{1}{n} \sum_{i=1}^{n} E(\hat{\theta}_n, \theta^*, X_i) \right)(\hat{\theta}_n - \theta^*)
\]
4
where $E(\hat{\theta}_n, \theta^*, X_i)$ is an error matrix that by Taylor’s theorem satisfies $\|E(\theta, \theta^*, x)\|_{op} \leq H(x)\|\theta - \theta^*\|$ for $\theta$ near $\theta^*$. As $\|\hat{\theta}_n - \theta^*\| = o_P(1)$, we obtain that

$$0 = \nabla \hat{R}_n(\hat{\theta}_n) = \nabla \hat{R}_n(\theta^*) + \left( \nabla^2 \hat{R}_n(\theta^*) + o_P(1) \right) (\hat{\theta}_n - \theta^*)$$

(1)

as

$$\left\| n^{-1} \sum_{i=1}^{n} E(\hat{\theta}_n, \theta^*, X_i) \right\|_{op} \leq \frac{1}{n} \sum_{i=1}^{n} H(X_i)\|\hat{\theta}_n - \theta^*\| = O_P(1) \cdot o_P(1) = o_P(1)$$

because $\frac{1}{n} \sum_{i=1}^{n} H(X_i) \overset{a.s.}{\rightarrow} \mathbb{E}[H(X)]$. In particular, rearranging the equality (1) we have

$$0 = \nabla \hat{R}_n(\theta^*) + \left( \nabla^2 \hat{R}_n(\theta^*) + o_P(1) \right) (\hat{\theta}_n - \theta^*).$$

We have $\nabla^2 \hat{R}_n(\theta^*) + o_P(1) \overset{P}{\rightarrow} \nabla^2 R(\theta^*)$, and it is eventually invertible, so

$$\hat{\theta}_n - \theta^* = -\left( \nabla^2 \hat{R}_n(\theta^*) + o_P(1) \right)^{-1} \nabla \hat{R}_n(\theta^*).$$

Multiplying by $\sqrt{n}$ and applying Slutsky’s theorem gives the result.

\[ \square \]