1. **(Miscellaneous short questions.)**

   (a) Let $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a convex loss, and fix any distribution on (x, y); consider our familiar setting of risk minimization for linear functions, meaning $f(w) := \mathbb{E}\ell(\langle w, -xy \rangle)$. Show that given a random draw (x, y) and any $g \in \partial\ell(\langle w, -xy \rangle)$, then $\mathbb{E}(-xyg) \in \partial f(w)$.
   Remark: this problem justifies the choice of stochastic gradient descent used in practice.
   Recall: the subgradient ∂h is defined as

$$\partial h(w) = \left\{ s \in \mathbb{R}^d \; : \; \forall v \in \mathbb{R}^d \centerdot h(v) \geq h(w) + \langle s, v - w \rangle \right\}.$$

   (b) Suppose $\Phi : \mathbb{R}^d \to \mathbb{R}$ is λ-strongly-convex (λ-sc) and differentiable, and define the *Bregman divergence*

$$D_\Phi(x, y) := \Phi(x) - \Big(\Phi(y) + \langle \nabla\Phi(y), x - y \rangle \Big).$$

   Prove that D_Φ is λ-sc in its first argument.
   (**Remark.** What about the second argument? Does a weaker property hold?)

   (c) Once again let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be λ-sc. Recall the definition of *Fenchel conjugate* $\Phi^*(s) := \sup_{x \in \mathbb{R}^d} \langle x, s \rangle - \Phi(s)$.
   The update rule of mirror descent may be written

$$w' := \arg\min_v \eta \langle \nabla f(w), v \rangle + D_\Phi(v, w).$$

   Prove this is equivalent to

$$w'' := \nabla\Phi^* \left(\Phi(w) - \eta \nabla f(w) \right).$$

   Hint: since Φ is strongly convex, then $(\nabla\Phi)^{-1}$ exists and is equal to $\nabla\Phi^*$ (you may use this without proof).

   (d) Suppose $Q \in \mathbb{R}^{d \times d}$ is symmetric positive definite, let $b \in \mathbb{R}^d$ be arbitary, and define $f(x) := \frac{1}{2}x^\top Qx + b^\top x$. Using direct computation (and not the preceding inverse gradient gradient fact), derive the Fenchel conjugate f^*, and prove it is correct.

   (e) Now suppose $Q \in \mathbb{R}^{d \times d}$ is merely symmetric positive *semi-definite* (it may fail to have an inverse), $b \in \mathbb{R}^d$ is again arbitrary, and define $f(x) := \frac{1}{2}x^\top Qx + b^\top x$. Derive the Fenchel conjugate f^*, and prove it is correct.

   (f) Freedman's inequality (Bernstein's inequality for martingales) implies: given martingale difference sequence $(Z_i)_{i=1}^n$ with $|Z_i| \leq b$ and $\sum_i \mathbb{E}(Z_i^2 | Z_{<i}) \leq v$, then with probability at least $1 - \delta$,

$$\sum_i Z_i \leq \sqrt{2v \ln(1/\delta)} + \frac{b \ln(1/\delta)}{3}.$$

   Consider the setting of the theorem in Lecture 15, but additionally $\mathbb{E}(g_i^2 | w_{i-1}) \leq \sigma^2$, and that for any given w_{i-1} it is possible to obtain an arbitrary number of mutually conditionally independent stochastic gradients g_i with all stated properties.
   Use all these assumptions together with the above version of Freedman's inequality to provide a refinement of the theorem in Lecture 15.

   (g) Consider the setting of the previous part, but suppose a minibatch of size b is used (b conditionally independent stochastic gradients are averaged together for each step). State the optimal values of step size η and batch size b by optimizing the right hand side of the previous bound.

Solution.

(Your solution here.)

1. **(Miscellaneous short questions.)**

   (a) Let $\ell : \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a convex loss, and fix any distribution on (x, y); consider our familiar setting of risk minimization for linear functions, meaning $f(w) := \mathbb{E}\ell(\langle w, -xy \rangle)$. Show that given a random draw (x, y) and any $g \in \partial\ell(\langle w, -xy \rangle)$, then $\mathbb{E}(-xyg) \in \partial f(w)$.
   Remark: this problem justifies the choice of stochastic gradient descent used in practice.
   Recall: the subgradient ∂h is defined as

$$\partial h(w) = \left\{ s \in \mathbb{R}^d \; : \; \forall v \in \mathbb{R}^d \centerdot h(v) \geq h(w) + \langle s, v - w \rangle \right\}.$$

   (b) Suppose $\Phi : \mathbb{R}^d \to \mathbb{R}$ is λ-strongly-convex (λ-sc) and differentiable, and define the *Bregman divergence*

$$D_\Phi(x, y) := \Phi(x) - \Big(\Phi(y) + \langle \nabla\Phi(y), x - y \rangle \Big).$$

   Prove that D_Φ is λ-sc in its first argument.
   (**Remark.** What about the second argument? Does a weaker property hold?)

   (c) Once again let $\Phi : \mathbb{R}^d \to \mathbb{R}$ be λ-sc. Recall the definition of *Fenchel conjugate* $\Phi^*(s) := \sup_{x \in \mathbb{R}^d} \langle x, s \rangle - \Phi(s)$.
   The update rule of mirror descent may be written

$$w' := \arg\min_v \eta \langle \nabla f(w), v \rangle + D_\Phi(v, w).$$

   Prove this is equivalent to

$$w'' := \nabla\Phi^* \left(\nabla\Phi(w) - \eta \nabla f(w) \right).$$

   Hint: since Φ is strongly convex, then $(\nabla\Phi)^{-1}$ exists and is equal to $\nabla\Phi^*$ (you may use this without proof).

   (d) Suppose $Q \in \mathbb{R}^{d \times d}$ is symmetric positive definite, let $b \in \mathbb{R}^d$ be arbitary, and define $f(x) := \frac{1}{2}x^\top Qx + b^\top x$. Using direct computation (and not the preceding inverse gradient gradient fact), derive the Fenchel conjugate f^*, and prove it is correct.

   (e) Now suppose $Q \in \mathbb{R}^{d \times d}$ is merely symmetric positive *semi-definite* (it may fail to have an inverse), $b \in \mathbb{R}^d$ is again arbitrary, and define $f(x) := \frac{1}{2}x^\top Qx + b^\top x$. Derive the Fenchel conjugate f^*, and prove it is correct.

   (f) Freedman's inequality (Bernstein's inequality for martingales) implies: given martingale difference sequence $(Z_i)_{i=1}^n$ with $|Z_i| \leq b$ and $\sum_i \mathbb{E}(Z_i^2 | Z_{<i}) \leq v$, then with probability at least $1 - \delta$,

$$\sum_i Z_i \leq \sqrt{2v \ln(1/\delta)} + \frac{b \ln(1/\delta)}{3}.$$

   Consider the setting of the theorem in Lecture 15, but additionally $\mathbb{E}(\|g_i - s_i\|^2 \mid w_{i-1}) \leq \sigma^2$, and that for any given w_{i-1} it is possible to obtain an arbitrary number of mutually conditionally independent stochastic gradients g_i with all stated properties.

   Use all these assumptions together with the above version of Freedman's inequality to provide a refinement of the theorem in Lecture 15.

   (g) Consider the setting of the previous part, but suppose a minibatch of size b is used (b conditionally independent stochastic gradients are averaged together for each step). State the optimal values of step size η and batch size b by optimizing the right hand side of the previous bound.

Solution.

(Your solution here.)

2