A. Online-to-batch conversion

Let $H$ be a finite hypothesis set of functions mapping from $X$ to $\mathbb{R}$ and $\ell : \mathbb{R} \times \mathbb{Y} \to \mathbb{R}_+$ a convex function bounded by $M$, convex with respect to its first argument. Let $A$ be an online learning algorithm that at each round returns a probability distribution $p_t$ over $H$. The goal of this problem is to study an online-to-batch conversion from these probability distributions into a randomized algorithm.

Let $P$ be the set of suffixes $P_t = \{p_t, \ldots, p_T\}$, $t = 1, \ldots, T$. Fix $\delta > 0$. For each $P \in P$, we define:

$$
\Gamma(P) = \frac{1}{|P|} \sum_{p_t \in P} \sum_{h \in H} p_t(h) \ell(h(x_t), y_t) + M \sqrt{\frac{\log T}{|P|}}.
$$

The online-to-batch conversion is done in two steps: first, a distribution $P_\delta$ is selected via $P_\delta \in \arg\min_{P \in P} \Gamma(P)$; next, a randomized algorithm is defined via the distribution $p$ over $H$ defined for any $h \in H$ by:

$$
p(h) = \frac{1}{|P_\delta|} \sum_{p_t \in P_\delta} p_t(h).
$$

Let $h_{\text{rand}}$ be the randomized hypothesis thereby defined.

1. Show that for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample $S = ((x_1, y_1), \ldots, (x_T, y_T))$ from $D$, the following inequality holds:

$$
\mathbb{E}[\ell(h_{\text{rand}}(x, y))] \leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{h \sim p_t} [\ell(h(x_t), y_t)] + M \sqrt{\frac{\log T}{T}}.
$$

Hint: you can apply Azuma’s inequality to an appropriately chosen martingale sequence.
Solution: Let \( \mathcal{P} = \{ p_{t_1}, \ldots, p_{t_\mathcal{P}} \} \) and let \( h_\mathcal{P} \) be the randomized hypothesis defined by the distribution \( p_\mathcal{P}(h) = \frac{1}{|\mathcal{P}|} \sum_{s=1}^{|\mathcal{P}|} p_{t_s}(h) \). Then,

\[
\mathbb{E}[\ell(h_\mathcal{P}(x, y))] - \frac{1}{|\mathcal{P}|} \sum_{s=1}^{|\mathcal{P}|} \mathbb{E}[\ell(h(x_{t_s}), y_{t_s})] = \sum_{s=1}^{|\mathcal{P}|} \sum_{h \in \mathcal{H}} \frac{p_{t_s}(h)}{|\mathcal{P}|} \left[ \mathbb{E}[\ell(h(x), y)] - \ell(h(x_{t_s}), y_{t_s}) \right].
\]

Let \( A_s \) denote the random variable \( \sum_{h \in \mathcal{H}} \frac{p_{t_s}(h)}{|\mathcal{P}|} \left[ \mathbb{E}[\ell(h(x), y)] - \ell(h(x_{t_s}), y_{t_s}) \right] \). Then, \( A_s \) forms a martingale sequence with respect to the filtration \( \mathcal{F}_{t_s} \), where \( \mathcal{F}_t \) is the \( \sigma \)-algebra generated by \( ((x_1, y_1), \ldots, (x_t, y_t)) \) since:

\[
\mathbb{E}[A_s | \mathcal{F}_{t_s}] = \frac{1}{|\mathcal{P}|} \sum_{h \in \mathcal{H}} \mathbb{E}[p_{t_s}(h) \mathbb{E}[\ell(h(x), y)] | \mathcal{F}_{t_s}] - \mathbb{E}[p_{t_s}(h) \ell(h(x_{t_s}), y_{t_s}) | \mathcal{F}_{t_s}],
\]

and, since \( p_t \) is completely determined by \( \mathcal{F}_{t-1} \) and \( (x_t, y_t) \) is independent of \( \mathcal{F}_{t-1} \), we have

\[
\mathbb{E}[p_{t_s}(h) \ell(h(x_{t_s}), y_{t_s}) | \mathcal{F}_{t_s}] = \mathbb{E}_{(x_{t_s-1}, y_{t_s-1})} \left[ \mathbb{E}_{(x_{t_s}, y_{t_s})} [p_{t_s}(h) \ell(h(x_{t_s}), y_{t_s}) | \mathcal{F}_{t_s}] \right] = \mathbb{E}_{(x_{t_s-1}, y_{t_s-1})} [p_{t_s}(h) \mathbb{E}_{(x_{t_s}, y_{t_s})} [\ell(h(x_{t_s}), y_{t_s}) | \mathcal{F}_{t_s}]].
\]

Thus, \( \mathbb{E}[A_s | \mathcal{F}_{t_s}] = 0 \). Therefore, by Azuma’s inequality, since \( |A_s| \leq \frac{M}{|\mathcal{P}|} \), for any \( \delta > 0 \), with probability at least \( 1 - \delta \),

\[
\mathbb{E}[\ell(h_\mathcal{P}(x, y))] \leq \frac{1}{|\mathcal{P}|} \sum_{s=1}^{|\mathcal{P}|} \mathbb{E}[\ell(h(x_{t_s}), y_{t_s})] + M \sqrt{\frac{\log \frac{1}{\delta}}{|\mathcal{P}|}} = \Gamma(\mathcal{P}).
\]

By the union bound, for any \( \delta > 0 \), with probability at least \( 1 - \delta \), for any \( \mathcal{P} \),

\[
\mathbb{E}[\ell(h_\mathcal{P}(x, y))] \leq \frac{1}{|\mathcal{P}|} \sum_{s=1}^{|\mathcal{P}|} \mathbb{E}[\ell(h(x_{t_s}), y_{t_s})] + M \sqrt{\frac{\log \frac{T}{\delta}}{|\mathcal{P}|}} = \Gamma(\mathcal{P}).
\]

Thus,

\[
\mathbb{E}[\ell(h_{\text{rand}}(x, y))] \leq \frac{1}{|\mathcal{P}_\delta|} \sum_{s=1}^{|\mathcal{P}_\delta|} \mathbb{E}[\ell(h(x_{t_s}), y_{t_s})] + M \sqrt{\frac{\log \frac{T}{\delta}}{|\mathcal{P}_\delta|}} \leq \frac{1}{T} \sum_{s=1}^T \mathbb{E}[\ell(h(x_t), y_t)] + M \sqrt{\frac{\log \frac{T}{\delta}}{T}},
\]

2
since $\mathcal{P}$ contains $\{p_1, \ldots, p_T\}$, and $\mathcal{P}_\delta$ is a minimizer of $\Gamma(\mathcal{P})$ over all $\mathcal{P}$, including $\{p_1, \ldots, p_T\}$.

2. Let $R_T$ denote the expected regret of the online algorithm $A$. Then, show that for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of an i.i.d. sample $S = ((x_1, y_1), \ldots, (x_T, y_T))$ from $\mathcal{D}$, the following inequality holds:

$$\mathbb{E}[\ell(h_{\text{rand}}(x, y))] \leq \inf_{h \in \mathcal{H}} \mathbb{E}[\ell(h(x), y)] + \frac{R_T}{T} + 2M \sqrt{\frac{\log \frac{2T}{\delta}}{T}}.$$

**Solution:** Let $h^* \in \mathcal{H}$ be the minimizer of $\mathbb{E}[\ell(h(x), y)]$. By Hoeffding’s inequality,

$$\mathbb{P} \left[ \frac{1}{T} \sum_{t=1}^T \ell(h^*(x_t), y_t) - \mathbb{E}[\ell(h^*(x, y))] > M \sqrt{\frac{\log \frac{2}{\delta}}{T}} \right] \leq \frac{\delta}{2}.
$$

Combining this with the result of the previous question, by the union bound, with probability at least $1 - \delta$

$$\mathbb{E}[\ell(h_{\text{rand}}(x, y))] - \inf_{h \in \mathcal{H}} \mathbb{E}[\ell(h(x), y)]

\leq \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{h \sim p_t} \left[ \ell(h(x_t), y_t) \right] + M \sqrt{\frac{\log \frac{2}{\delta}}{T}} - \frac{1}{T} \sum_{t=1}^T \ell(h^*(x_t), y_t) + M \sqrt{\frac{\log \frac{2}{\delta}}{T}}

\leq \frac{R_T}{T} + 2M \sqrt{\frac{\log \frac{2T}{\delta}}{T}}.$$

**B. Mirror Descent**

The notation and definitions used are those adopted in lectures.

1. Prove that Mirror Descent coincides with EG when the convex set is the simplex and the unnormalized relative entropy is used as a Bregman divergence. In particular, you should show that the corresponding mirror map $\Phi$ is 1-strongly convex with respect to $\| \cdot \|_1$ on the simplex.

**Solution:** Use Pinsker’s inequality to show the 1-strong convexity.
2. Consider the scenario where the functions $f_t$ are differentiable and where, when requesting the gradient $\nabla f_t(w)$ of $f_t$ at $w$, the learner receives only a random variable $g_t(w)$, such that $E[g_t(w)] = \nabla f_t(w)$. When $w_t$ itself is a random variable, we have $E[g_t(w_t)|w_t] = \nabla f_t(w_t)$.

Show that MD in this scenario benefits from the following guarantee:

$$E[R_T(MD)] \leq \frac{B(w^* \parallel w_1)}{\eta} + \frac{\eta E[\|g_t(w_t)\|^2_2]}{2\alpha},$$

and that for an appropriate choice of $\eta$, we have

$$E[R_T(MD)] \leq DG^* \sqrt{\frac{2T}{\alpha}},$$

when $B(w^* \parallel w_1) \leq D^2$ and $E[\|g_t(w_t)\|^2_2] \leq G^*_2$.

**Solution:** Proceeding as in the proof for MD in the standard case and
taking expectations, we have:
\[
E[R_T(MD)] = \mathbb{E} \left\{ \sum_{t=1}^{T} \left( f_t(w_t) - f_t(w^*) \right) \right\} \\
\leq \mathbb{E} \left\{ \sum_{t=1}^{T} \mathbb{E} [g_t(w_t)w_t] \cdot (w_t - w^*) \right\} \quad \text{(def. of grad.)} \\
\leq \mathbb{E} \left\{ \sum_{t=1}^{T} \nabla f_t(w_t) \cdot (w_t - w^*) \right\} \quad \text{(tower rule)} \\
= \mathbb{E} \left\{ \frac{1}{\eta} \sum_{t=1}^{T} [\nabla \Phi(w_t) - \nabla \Phi(v_{t+1})] \cdot (w_t - w^*) \right\} \quad \text{(def. of v_t)} \\
= \frac{1}{\eta} \sum_{t=1}^{T} \left[ B(w^* \parallel w_t) - B(w^* \parallel v_{t+1}) + B(w_t \parallel v_{t+1}) \right] \quad \text{(Breg. div. Identity)} \\
\leq \mathbb{E} \left\{ \frac{1}{\eta} \sum_{t=1}^{T} \left[ B(w^* \parallel w_t) - B(w^* \parallel w_{t+1}) - B(w_{t+1} \parallel v_{t+1}) + B(w_t \parallel v_{t+1}) \right] \right\} \quad \text{(Pythagorean ineq.)} \\
= \frac{1}{\eta} \left[ B(w^* \parallel w_1) - B(w^* \parallel w_{T+1}) \right] + \mathbb{E} \left\{ \frac{1}{\eta} \sum_{t=1}^{T} \left[ - B(w_{t+1} \parallel v_{t+1}) + B(w_t \parallel v_{t+1}) \right] \right\} \quad \text{(telescoping sum)} \\
\leq \frac{B(w^* \parallel w_1)}{\eta} + \mathbb{E} \left\{ \frac{1}{\eta} \sum_{t=1}^{T} \left[ B(w_t \parallel v_{t+1}) - B(w_{t+1} \parallel v_{t+1}) \right] \right\} \quad \text{(non-negativity of Bregman div.)}
\[
\mathbb{E}[B(w_t \parallel v_{t+1})] - B(w_{t+1} \parallel v_{t+1})] \\
= \mathbb{E}[\Phi(w_t) - \Phi(w_{t+1}) - \nabla \Phi(v_{t+1}) \cdot (w_t - w_{t+1})] \\
\leq \mathbb{E}\left[\left(\nabla \Phi(w_t) - \nabla \Phi(v_{t+1})\right) \cdot (w_t - w_{t+1}) - \frac{\alpha}{2} \|w_t - w_{t+1}\|^2\right] \\
\leq \mathbb{E}\left[ -\eta g_t(w_t) \cdot (w_t - w_{t+1}) - \frac{\alpha}{2} \|w_t - w_{t+1}\|^2\right] \quad \text{(\(\alpha\)-strong convexity)} \\
= \mathbb{E}\left[ -\eta g_t(w_t) \cdot (w_t - w_{t+1}) - \frac{\alpha}{2} \|w_t - w_{t+1}\|^2\right] \quad \text{(def. of} v_{t+1}) \\
\leq \mathbb{E}\left[ \eta \|g_t(w_t)\|_* \|w_t - w_{t+1}\| - \frac{\alpha}{2} \|w_t - w_{t+1}\|^2\right] \\
\leq \eta^2 \mathbb{E}\left[\|g_t(w_t)\|^2\right] \quad \frac{1}{2\alpha}. \quad \text{(max. of 2nd deg. eq.)}
\]

3. In this question, we adopt the same assumptions as in the previous one except: the functions \(f_t\) are all equal to \(f\), \(f\) is assumed to be convex and \(\beta\)-smooth with respect to \(\|\cdot\|\) and, instead of the upper bound \(\mathbb{E}\left[\|g(w_t)\|^2\right] \leq G^2\), we will assume that the following bound on the variance holds for all \(w\): 
\[
\mathbb{E}\left[\|\nabla f(w) - g(w)\|^2\right] \leq \sigma^2.
\]

(a) Show that the following inequality holds:
\[
\sum_{t=1}^{T} f(w_{t+1}) - f(w^*) \\
\leq \sum_{t=1}^{T} \nabla f(w_t) \cdot (w_{t+1} - w_t) + \frac{\beta}{2} \|w_{t+1} - w_t\|^2 + \nabla f(w_t) \cdot (w_t - w^*).
\]

(b) Prove the identity \(2u \cdot v \leq \mu\|u\|^2 + \|v\|^2 / \mu\) valid for any \(\mu > 0\)
and vectors \( u \) and \( v \). Use that to show the following:

\[
\sum_{t=1}^{T} f(w_{t+1}) - f(w^*) \\
\leq \sum_{t=1}^{T} g(w_t) \cdot (w_{t+1} - w_t) + \nabla f(w_t) \cdot (w_t - w^*) \\
+ \sum_{t=1}^{T} \eta \left[ \left\| \nabla f(w_t) - g(w_t) \right\|_2^2 \right] + \frac{\beta + 1/\eta}{2} \| w_{t+1} - w_t \|^2.
\]

Solution: For the Cauchy-Schwarz-type inequality, observe that:

\[
0 \leq \left( \sqrt{\mu} \| u \|_* - \frac{1}{\sqrt{\mu}} \| v \| \right)^2 \\
= \mu \| u \|_*^2 + \frac{1}{\mu} \| v \|^2 - 2 \| u \|_* \| v \| \\
\leq \mu \| u \|_*^2 + \frac{1}{\mu} \| v \|^2 - 2 u \cdot v. \quad \text{(Hölder's ineq.)}
\]

(c) Use the 1-strong convexity of \( \Phi \) to show the following:

\[
\sum_{t=1}^{T} f(w_{t+1}) - f(w^*) \\
\leq \sum_{t=1}^{T} g(w_t) \cdot (w_{t+1} - w_t) + [\nabla f(w_t) - g(w_t)] \cdot (w_t - w^*) \\
+ \sum_{t=1}^{T} \frac{\eta}{2} \left[ \left\| \nabla f(w_t) - g(w_t) \right\|_*^2 \right] + (\beta + 1/\eta) B(w_{t+1} \parallel w_t).
\]

(d) Prove the following inequality:

\[
[\nabla \Phi(w_{t+1}) - \nabla \Phi(v_{t+1})] \cdot (w_{t+1} - w^*) \leq 0.
\]

(e) Use the previous question to prove:

\[
\frac{1}{\beta + 1/\eta} g(w_t) \cdot (w_{t+1} - w^*) \leq B(w^* \parallel w_t) - B(w^* \parallel w_{t+1}) - B(w_{t+1} \parallel w_t).
\]
(f) Use the previous results to conclude that the following regret bound holds for MD run with step size $\frac{1}{\beta + 1/\eta}$, with $\eta = \frac{D}{\sigma} \sqrt{\frac{2}{T}}$:

$$\mathbb{E}[R_T(MD)] \leq \beta D^2 + \sigma D \sqrt{2T}.$$ 

Solution: Proceeding as in the proof for MD in the standard case and
taking expectations, we have:

\[ E[R_T(MD)] = E \left[ \sum_{t=1}^{T} (f(w_{t+1}) - f(w)) \right] \]

\[ = E \left[ \sum_{t=1}^{T} (f(w_{t+1}) - f(w_t) + f(w_t) - f(w^*)) \right] \]

\[ \leq E \left[ \sum_{t=1}^{T} \nabla f(w_t) \cdot (w_{t+1} - w_t) + \frac{\beta}{2} \|w_{t+1} - w_t\|^2 + \nabla f(w_t) \cdot (w_t - w^*) \right] \]

\[ = E \left[ \sum_{t=1}^{T} g(w_t) \cdot (w_{t+1} - w_t) + \nabla f(w_t) \cdot (w_t - w^*) \right] \]

\[ + E \left[ \sum_{t=1}^{T} \nabla f(w_t) - g(w_t) \right] \cdot (w_{t+1} - w_t) + \frac{\beta}{2} \|w_{t+1} - w_t\|^2 \]

\[ = E \left[ \sum_{t=1}^{T} g(w_t) \cdot (w_{t+1} - w_t) + \nabla f(w_t) \cdot (w_t - w^*) \right] \]

\[ + E \left[ \sum_{t=1}^{T} \eta \left[ \|\nabla f(w_t) - g(w_t)\|^2 + \frac{\beta + 1/\eta}{2} \|w_{t+1} - w_t\|^2 \right] \right] \]

Cauchy-Schwarz-type inequality

\[ \leq E \left[ \sum_{t=1}^{T} g(w_t) \cdot (w_{t+1} - w_t) + \nabla f(w_t) \cdot (w_t - w^*) \right] \]

\[ + E \left[ \sum_{t=1}^{T} \eta \left[ \|\nabla f(w_t) - g(w_t)\|^2 + \frac{\beta + 1/\eta}{2} \|w_{t+1} - w_t\|^2 \right] \right] \]

(1-strong convexity of \( \Phi \))

\[ \leq E \left[ \sum_{t=1}^{T} g(w_t) \cdot (w_{t+1} - w^*) + \nabla f(w_t) \cdot (w_t - w^*) \right] \]

\[ + E \left[ \sum_{t=1}^{T} \eta \left[ \|\nabla f(w_t) - g(w_t)\|^2 + \frac{\beta + 1/\eta}{2} \|w_{t+1} - w_t\|^2 \right] \right] \]

\[ = E \left[ \sum_{t=1}^{T} g(w_t) \cdot (w_{t+1} - w^*) + \eta \left[ \|\nabla f(w_t) - g(w_t)\|^2 + \frac{\beta + 1/\eta}{2} \|w_{t+1} - w_t\|^2 \right] \right]. \]
Now, first, observe that:
\[
\nabla \Phi(w_{t+1}) - \nabla \Phi(v_{t+1}) \cdot (w_{t+1} - w^*)
= B(w^* \parallel w_{t+1}) + B(v_{t+1} \parallel v_{t+1}) - B(w^* \parallel v_{t+1})
\]
(Bregman div. identity)
\[
= B(w^* \parallel v_{t+1}) - B(w^* \parallel v_{t+1})
\]
(Pythagorean theorem)
\[
\leq 0.
\]

In view of that, we can write:
\[
\frac{1}{\beta + \frac{1}{\eta}} g(w_t) \cdot (w_{t+1} - w^*)
= [\nabla \Phi(w_t) - \nabla \Phi(v_{t+1})] \cdot (w_{t+1} - w^*)
\]
(by def. of MD update)
\[
= [\nabla \Phi(w_t) - \nabla \Phi(w_{t+1}) + \nabla \Phi(w_{t+1}) - \nabla \Phi(v_{t+1})] \cdot (w_{t+1} - w^*)
\]
\[
\leq [\nabla \Phi(w_t) - \nabla \Phi(w_{t+1})] \cdot (w_{t+1} - w^*)
= B(w^* \parallel w_t) - B(w^* \parallel w_{t+1}) - B(w_{t+1} \parallel w_t).
\]
(Bregman div. identity)

Thus, we have:
\[
\mathbb{E}[RT(MD)]
\leq \beta + \frac{\eta}{\eta} \frac{1}{\eta} g(w_t) \cdot (w_{t+1} - w^*)
= \frac{\eta}{2} \|\nabla f(w_t) - g(w_t)\|^2 + \frac{\eta}{2} \|\nabla f(w_t) - g(w_t)\|^2
\]
non-negativity of Bregman div.
\[
= \beta D^2 + \sigma D \sqrt{2T}.
\]