Mehryar Mohri
Advanced Machine Learning 2021
Courant Institute of Mathematical Sciences
Homework assignment 1
March 09, 2021
Due: March 23, 2021

A. Online-to-batch conversion

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

Let $\mathcal{P}$ be the set of suffixes $\mathcal{P}_t$: $\mathcal{P}_t = \{\mathsf{p}_t, \ldots, \mathsf{p}_T\}$, $t = 1, \ldots, T$. Fix $\delta > 0$. For each $\mathcal{P} \in \mathcal{P}$, we define:

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

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

$$\mathsf{p}(h) = \frac{1}{|\mathcal{P}_\delta|} \sum_{\mathsf{p}_t \in \mathcal{P}_\delta} \mathsf{p}_t(h).$$

Let h_{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 $\mathcal{D}$, the following inequality holds:

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

Hint: you can apply Azuma's inequality to an appropriately chosen martingale sequence.

Solution: Let $\mathcal{P} = \{\mathsf{p}_{t_1}, \ldots, \mathsf{p}_{t_{|\mathcal{P}|}}\}$ and let $h_{\mathcal{P}}$ be the randomized hypothesis defined by the distribution $\mathsf{p}_{\mathcal{P}}(h) = \frac{1}{|\mathcal{P}|} \sum_{s=1}^{|\mathcal{P}|} \mathsf{p}_{t_s}(h)$. Then,

$$\begin{aligned}
&\mathbb{E}[\ell(h_{\mathcal{P}}(x,y))] - \frac{1}{|\mathcal{P}|} \sum_{s=1}^{|\mathcal{P}|} \mathop{\mathbb{E}}_{h \sim \mathsf{p}_t} [\ell(h(x_{t_s}), y_{t_s})] \\
&= \sum_{s=1}^{|\mathcal{P}|} \sum_{h \in \mathcal{H}} \frac{\mathsf{p}_{t_s}(h)}{|\mathcal{P}|} \Big[\mathbb{E}[\ell(h(x), y)] - \ell(h(x_{t_s}), y_{t_s}) \Big].
\end{aligned}$$

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

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

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

$$\begin{aligned}
\mathbb{E}[\mathsf{p}_{t_s}(h) \ell(h(x_{t_s}), y_{t_s}) | \mathcal{F}_{t_s}] &= \mathop{\mathbb{E}}_{(x_1^{t_s - 1}, y_1^{t_s - 1})} [\mathop{\mathbb{E}}_{(x_{t_s}, y_{t_s})} [\mathsf{p}_{t_s}(h) \ell(h(x_{t_s}), y_{t_s}) | \mathcal{F}_{t_s}]] \\
&= \mathop{\mathbb{E}}_{(x_1^{t_s - 1}, y_1^{t_s - 1})} [\mathsf{p}_{t_s}(h) \mathop{\mathbb{E}}_{(x_{t_s}, y_{t_s})} [\ell(h(x_{t_s}), y_{t_s}) | \mathcal{F}_{t_s}]].
\end{aligned}$$

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 $1 - \delta$,

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

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

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

Thus,

$$\begin{aligned}
\mathbb{E}[\ell(h_{\text{rand}}(x,y))] &\leq \frac{1}{|\mathcal{P}_\delta|} \sum_{s=1}^{|\mathcal{P}_\delta|} \mathop{\mathbb{E}}_{h \sim \mathsf{p}_t} [\ell(h(x_{t_s}), y_{t_s})] + M \sqrt{\frac{\log \frac{T}{\delta}}{|\mathcal{P}_\delta|}} \\
&\leq \frac{1}{T} \sum_{s=1}^{T} \mathop{\mathbb{E}}_{h \sim \mathsf{p}_t} [\ell(h(x_t), y_t)] + M \sqrt{\frac{\log \frac{T}{\delta}}{T}},
\end{aligned}$$

since $\mathcal{P}$ contains $\{\mathsf{p}_1, \ldots, \mathsf{p}_T\}$, and $\mathcal{P}_\delta$ is a minimizer of $\Gamma(\mathcal{P})$ over all $\mathcal{P}$, including $\{\mathsf{p}_1, \ldots, \mathsf{p}_T\}$.

2. Let R_T denote the expected regret of the online algorithm $\mathcal{A}$. 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 $1 - \delta$

$$\begin{aligned}
&\mathbb{E}[\ell(h_{\text{rand}}(x, y))] - \inf_{h \in \mathcal{H}} \mathbb{E}[\ell(h(x), y)] \\
&\leq \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{h \sim \mathsf{p}_t}[\ell(h(x_t), y_t)] + M\sqrt{\frac{\log \frac{2T}{\delta}}{T}} - \frac{1}{T}\sum_{t=1}^{T} \ell(h^*(x_t), y_t)) + M\sqrt{\frac{\log \frac{2}{\delta}}{T}} \\
&\leq \frac{R_T}{T} + 2M\sqrt{\frac{\log \frac{2T}{\delta}}{T}}.
\end{aligned}$$

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 Φ 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 $\nabla f_t(\mathbf{w})$ of f_t at $\mathbf{w}$, the learner receives only a random variable $g_t(\mathbf{w})$, such that $\mathbb{E}[g_t(\mathbf{w})] = \nabla f_t(\mathbf{w})$. When $\mathbf{w}_t$ itself is a random variable, we have $\mathbb{E}[g_t(\mathbf{w}_t)|\mathbf{w}_t] = \nabla f_t(\mathbf{w}_t)$. Show that MD in this scenario benefits from the following guarantee:

$$\mathbb{E}[R_T(\mathrm{MD})] \leq \frac{\mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_1)}{\eta} + \frac{\eta\,\mathbb{E}\big[\|g_t(\mathbf{w}_t)\|_*^2\big]}{2\alpha},$$

and that for an appropriate choice of η, we have

$$\mathbb{E}[R_T(\mathrm{MD})] \leq DG_* \sqrt{\frac{2T}{\alpha}},$$

when $\mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_1) \leq D^2$ and $\mathbb{E}\big[\|g_t(\mathbf{w}_t)\|_*^2\big] \leq G_*^2$.

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

$$
\begin{aligned}
&\mathbb{E}[R_T(\mathrm{MD})] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\big(f_t(\mathbf{w}_t) - f_t(\mathbf{w}^*)\big)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}[g_t(\mathbf{w}_t)|\mathbf{w}_t]\cdot(\mathbf{w}_t - \mathbf{w}^*)\right] && \text{(def. of grad.)} \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\nabla f_t(\mathbf{w}_t)\cdot(\mathbf{w}_t - \mathbf{w}^*)\right] && \text{(tower rule)} \\
&= \mathbb{E}\left[\frac{1}{\eta}\sum_{t=1}^{T}[\nabla\Phi(\mathbf{w}_t) - \nabla\Phi(\mathbf{v}_{t+1})]\cdot(\mathbf{w}_t - \mathbf{w}^*)\right] && \text{(def. of } \mathbf{v}_t\text{)} \\
&= \frac{1}{\eta}\sum_{t=1}^{T}\big[\mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_t) - \mathsf{B}(\mathbf{w}^* \parallel \mathbf{v}_{t+1}) + \mathsf{B}(\mathbf{w}_t \parallel \mathbf{v}_{t+1})\big] \\
& && \text{(Breg. div. Identity)} \\
&\leq \mathbb{E}\left[\frac{1}{\eta}\sum_{t=1}^{T}\big[\mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_t) - \mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_{t+1}) - \mathsf{B}(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1}) + \mathsf{B}(\mathbf{w}_t \parallel \mathbf{v}_{t+1})\big]\right] \\
& && \text{(Pythagorean ineq.)} \\
&= \frac{1}{\eta}\Big[\mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_1) - \mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_{T+1})\Big] + \mathbb{E}\left[\frac{1}{\eta}\sum_{t=1}^{T}\Big[- \mathsf{B}(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1}) + \mathsf{B}(\mathbf{w}_t \parallel \mathbf{v}_{t+1})\Big]\right] \\
& && \text{(telescoping sum)} \\
&\leq \frac{\mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_1)}{\eta} + \mathbb{E}\left[\frac{1}{\eta}\sum_{t=1}^{T}\Big[\mathsf{B}(\mathbf{w}_t \parallel \mathbf{v}_{t+1}) - \mathsf{B}(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1})\Big]\right]. \\
& && \text{(non-negativity of Bregman div.)}
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}[\mathsf{B}(\mathbf{w}_t \parallel \mathbf{v}_{t+1}) - \mathsf{B}(\mathbf{w}_{t+1} \parallel \mathbf{v}_{t+1})] \\
&= \mathbb{E}[\Phi(\mathbf{w}_t) - \Phi(\mathbf{w}_{t+1}) - \nabla\Phi(\mathbf{v}_{t+1}) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1})] \\
&\leq \mathbb{E}\Big[\big(\nabla\Phi(\mathbf{w}_t) - \nabla\Phi(\mathbf{v}_{t+1})\big) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) - \frac{\alpha}{2}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2\Big] \\
&\hspace{20em} (\alpha\text{-strong convexity}) \\
&= \mathbb{E}\Big[-\eta\, g_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) - \frac{\alpha}{2}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2\Big] \qquad (\text{def. of } \mathbf{v}_{t+1}) \\
&= \mathbb{E}\Big[-\eta\, g_t(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}_{t+1}) - \frac{\alpha}{2}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2\Big] \\
&\leq \mathbb{E}\Big[\eta\, \|g_t(\mathbf{w}_t)\|_* \,\|\mathbf{w}_t - \mathbf{w}_{t+1}\| - \frac{\alpha}{2}\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2\Big] \\
&\leq \frac{\eta^2\, \mathbb{E}\big[\|g_t(\mathbf{w}_t)\|_*^2\big]}{2\alpha}. \qquad (\text{max. of 2nd deg. eq.})
\end{aligned}
$$

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 β-smooth with respect to $\|\cdot\|$ and, instead of the upper bound $\mathbb{E}\big[\|g(\mathbf{w}_t)\|_*^2\big] \leq G_*^2$, we will assume that the following bound on the variance holds for all $\mathbf{w}$:

$$\mathbb{E}\big[\|\nabla f(\mathbf{w}) - g(\mathbf{w})\|_*^2\big] \leq \sigma^2.$$

   (a) Show that the following inequality holds:

$$
\begin{aligned}
&\sum_{t=1}^{T} f(\mathbf{w}_{t+1}) - f(\mathbf{w}^*) \\
&\leq \sum_{t=1}^{T} \nabla f(\mathbf{w}_t) \cdot (\mathbf{w}_{t+1} - \mathbf{w}_t) + \frac{\beta}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 + \nabla f(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*).
\end{aligned}
$$

   (b) Prove the identity $2\mathbf{u} \cdot \mathbf{v} \leq \mu\|\mathbf{u}\|_*^2 + \|\mathbf{v}\|^2/\mu$ valid for any $\mu > 0$

and vectors $\mathbf{u}$ and $\mathbf{v}$. Use that to show the following:

$$\begin{aligned}
&\sum_{t=1}^{T} f(\mathbf{w}_{t+1}) - f(\mathbf{w}^*) \\
&\leq \sum_{t=1}^{T} g(\mathbf{w}_t) \cdot (\mathbf{w}_{t+1} - \mathbf{w}_t) + \nabla f(\mathbf{w}_t) \cdot (\mathbf{w}_t - \mathbf{w}^*) \\
&+ \sum_{t=1}^{T} \frac{\eta}{2} \left[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2 \right] + \frac{\beta + 1/\eta}{2} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2.
\end{aligned}$$

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

$$\begin{aligned}
0 &\leq \left(\sqrt{\mu} \|\mathbf{u}\|_* - \frac{1}{\sqrt{\mu}} \|\mathbf{v}\| \right)^2 \\
&= \mu \|\mathbf{u}\|_*^2 + \frac{1}{\mu} \|\mathbf{v}\|^2 - 2 \|\mathbf{u}\|_* \|\mathbf{v}\| \\
&\leq \mu \|\mathbf{u}\|_*^2 + \frac{1}{\mu} \|\mathbf{v}\|^2 - 2 \mathbf{u} \cdot \mathbf{v}. \qquad \text{(Hölder's ineq.)}
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{t=1}^{T} f(\mathbf{w}_{t+1}) - f(\mathbf{w}^*) \\
&\leq \sum_{t=1}^{T} g(\mathbf{w}_t) \cdot (\mathbf{w}_{t+1} - \mathbf{w}^*) + [\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)] \cdot (\mathbf{w}_t - \mathbf{w}^*) \\
&+ \sum_{t=1}^{T} \frac{\eta}{2} \left[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2 \right] + (\beta + 1/\eta) \mathsf{B}(\mathbf{w}_{t+1} \parallel \mathbf{w}_t).
\end{aligned}$$

(d) Prove the following inequality:

$$[\nabla \Phi(\mathbf{w}_{t+1}) - \nabla \Phi(\mathbf{v}_{t+1})] \cdot (\mathbf{w}_{t+1} - \mathbf{w}^*) \leq 0.$$

(e) Use the previous question to prove:

$$\frac{1}{\beta + \frac{1}{\eta}} g(\mathbf{w}_t) \cdot (\mathbf{w}_{t+1} - \mathbf{w}^*) \leq \mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_t) - \mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_{t+1}) - \mathsf{B}(\mathbf{w}_{t+1} \parallel \mathbf{w}_t).$$

(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 $\eta = \frac{D}{\sigma}\sqrt{\frac{2}{T}}$:

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

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

$$
\begin{aligned}
&\mathbb{E}[R_T(\mathrm{MD})] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\big(f(\mathbf{w}_{t+1}) - f(\mathbf{w}^*)\big)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\big(f(\mathbf{w}_{t+1}) - f(\mathbf{w}_t) + f(\mathbf{w}_t) - f(\mathbf{w}^*)\big)\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\nabla f(\mathbf{w}_t)\cdot(\mathbf{w}_{t+1}-\mathbf{w}_t) + \frac{\beta}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2 + \nabla f(\mathbf{w}_t)\cdot(\mathbf{w}_t-\mathbf{w}^*)\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\beta\text{-smoothness and def. of grad.}) \\
&= \mathbb{E}\left[\sum_{t=1}^{T} g(\mathbf{w}_t)\cdot(\mathbf{w}_{t+1}-\mathbf{w}_t) + \nabla f(\mathbf{w}_t)\cdot(\mathbf{w}_t-\mathbf{w}^*)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}[\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)]\cdot(\mathbf{w}_{t+1}-\mathbf{w}_t) + \frac{\beta}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} g(\mathbf{w}_t)\cdot(\mathbf{w}_{t+1}-\mathbf{w}_t) + \nabla f(\mathbf{w}_t)\cdot(\mathbf{w}_t-\mathbf{w}^*)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta}{2}\big[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2\big] + \frac{\beta+1/\eta}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{Cauchy-Schwarz-type inequality}) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} g(\mathbf{w}_t)\cdot(\mathbf{w}_{t+1}-\mathbf{w}_t) + \nabla f(\mathbf{w}_t)\cdot(\mathbf{w}_t-\mathbf{w}^*)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta}{2}\big[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2\big] + (\beta+1/\eta)\mathsf{B}(\mathbf{w}_{t+1}\parallel\mathbf{w}_t)\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{1-strong convexity of }\Phi) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} g(\mathbf{w}_t)\cdot(\mathbf{w}_{t+1}-\mathbf{w}^*) + [\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)]\cdot(\mathbf{w}_t-\mathbf{w}^*)\right] \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\frac{\eta}{2}\big[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2\big] + (\beta+1/\eta)\mathsf{B}(\mathbf{w}_{t+1}\parallel\mathbf{w}_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} g(\mathbf{w}_t)\cdot(\mathbf{w}_{t+1}-\mathbf{w}^*) + \frac{\eta}{2}\big[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2\big] + (\beta+1/\eta)\mathsf{B}(\mathbf{w}_{t+1}\parallel\mathbf{w}_t)\right].
\end{aligned}
$$

Now, first, observe that:

$$\begin{aligned}
&[\nabla\Phi(\mathbf{w}_{t+1}) - \nabla\Phi(\mathbf{v}_{t+1})] \cdot (\mathbf{w}_{t+1} - \mathbf{w}^*) \\
&= \mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_{t+1}) + \mathsf{B}(\mathbf{w}_{t+1}) \parallel \mathbf{v}_{t+1}) - \mathsf{B}(\mathbf{w}^* \parallel \mathbf{v}_{t+1}) \\
&\hspace{18em}\text{(Bregman div. identity)} \\
&= \mathsf{B}(\mathbf{w}^* \parallel \mathbf{v}_{t+1}) - \mathsf{B}(\mathbf{w}^* \parallel \mathbf{v}_{t+1}) \qquad \text{(Pythagorean theorem)} \\
&\le 0.
\end{aligned}$$

In view of that, we can write:

$$\begin{aligned}
&\frac{1}{\beta + \frac{1}{\eta}} g(\mathbf{w}_t) \cdot (\mathbf{w}_{t+1} - \mathbf{w}^*) \\
&= [\nabla\Phi(\mathbf{w}_t) - \nabla\Phi(\mathbf{v}_{t+1})] \cdot (\mathbf{w}_{t+1} - \mathbf{w}^*) \qquad \text{(by def. of MD update)} \\
&= [\nabla\Phi(\mathbf{w}_t) - \nabla\Phi(\mathbf{w}_{t+1}) + \nabla\Phi(\mathbf{w}_{t+1}) - \nabla\Phi(\mathbf{v}_{t+1})] \cdot (\mathbf{w}_{t+1} - \mathbf{w}^*) \\
&\le [\nabla\Phi(\mathbf{w}_t) - \nabla\Phi(\mathbf{w}_{t+1})] \cdot (\mathbf{w}_{t+1} - \mathbf{w}^*) \\
&= \mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_t) - \mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_{t+1}) - \mathsf{B}(\mathbf{w}_{t+1} \parallel \mathbf{w}_t). \\
&\hspace{18em}\text{(Bregman div. identity)}
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
&\mathbb{E}[R_T(\text{MD})] \\
&\mathbb{E}\left[\sum_{t=1}^{T} g(\mathbf{w}_t) \cdot (\mathbf{w}_{t+1} - \mathbf{w}^*) + \frac{\eta}{2}\left[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2\right] + (\beta + 1/\eta)\mathsf{B}(\mathbf{w}_{t+1} \parallel \mathbf{w}_t)\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} (\beta + 1/\eta)[\mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_t) - \mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_{t+1})] + \frac{\eta}{2}\left[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2\right]\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} (\beta + 1/\eta)[\mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_t) - \mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_{t+1})] + \frac{\eta}{2}\left[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2\right]\right] \\
&= \mathbb{E}\left[(\beta + 1/\eta)[\mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_1) - \mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_{T+1})] + \frac{\eta T}{2}\left[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2\right]\right] \\
&\le \mathbb{E}\left[(\beta + 1/\eta)[\mathsf{B}(\mathbf{w}^* \parallel \mathbf{w}_1)] + \frac{\eta T}{2}\left[\|\nabla f(\mathbf{w}_t) - g(\mathbf{w}_t)\|_*^2\right]\right] \\
&\hspace{14em}\text{(non-negativity of Bregman div.)} \\
&\le (\beta + 1/\eta) D^2 + \frac{\eta T}{2}\sigma^2 \\
&= \beta D^2 + \sigma D\sqrt{2T}.
\end{aligned}$$