Machine Learning Theory (CS 6783)

Lecture 19: Burkholder Method

1 Recap

We saw in last class that whenever we have a relaxation $\mathbf{Rel}_n$ w.r.t some function $\phi : (\mathcal{X} \times \mathcal{Y})^n \mapsto \mathbb{R}$ that satisfies the conditions:
1. Dominance condition : $\mathbf{Rel}_n(x_{1:n}, y_{1:n}) \geq -\phi((x_1, y_1), \ldots, (x_n, y_n))$
2. Final condition : $\mathbf{Rel}_n(\cdot) \leq 0$
3. Admissibility condition : For any $x_1, \ldots, x_t \in \mathcal{X}$ and any $y_1, \ldots, y_{t-1} \in \mathcal{Y}$,

$$\inf_{q_t \in \Delta(\mathcal{Y})} \sup_{y_t \in \mathcal{Y}} \{\mathbb{E}_{\hat{y}_t \sim q_t}[\ell(\hat{y}_t, y_t)] + \mathbf{Rel}_n(x_{1:t}, y_{1:t})\} \leq \mathbf{Rel}_n(x_{1:t-1}, y_{1:t-1})$$

then one can obtain online learning algorithm using such a relaxation. The key idea is that if the relaxation is computationally easy to handle, then the algorithm it implies is efficient. Hence the name of the game is to find such nice relaxations. We will explore in this lecture some methods to find such nice relaxations in many cases that are of practical interest.

2 Sufficient Statistics

The way we set up relaxation has an additional disadvantage that we need to know n in advance. Another issue is that while the relaxations are powerful, often to obtain efficient algorithms we need to use convex losses and in these cases we would like to linearize the loss using gradient of the loss. Relaxations are not directly amenable to these modifications. Finally, one of the key advantages of online learning methods are their computational efficiency and fact that they only keep current version of some information about the predictor and keep updating it on every time step instead of carrying around all the past instances. However relaxation at some time step by the way its written seems to suggest that we keep around all past data. Throughout we shall consider the following online learning game:

For $t = 1$ to n

Adversary provides input instance x_t
Learner picks $\hat{y}_t \in \widehat{\mathcal{Y}}$
Adversary picks outcome $y_t \in \mathcal{Y}$ and learner suffer's loss $\ell(\hat{y}_t, y_t)$

End For

Now we introduce the concept of additive sufficient statistics for online learning. The key idea is that we want to develop algorithms that don't need to store the entire sequence of instances but rather only some aggregate information about the past that we update one every round.

Definition 1. *Let $\mathcal{T}$ be some vector space. A function $\mathbf{T} : \mathcal{X} \times \mathbb{R} \mapsto \mathcal{T}$ is an additive sufficient statistic for ϕ if there exists some $V : \mathcal{T} \mapsto \mathbb{R}$ s.t. for any n,*

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \phi(x_1, \hat{y}_1, y_1, \ldots, x_n, , \hat{y}_n, y_n) \leq \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + V\left(\sum_{t=1}^{n} \mathbf{T}\left(x_t, \partial\ell(\hat{y}_t, y_t)\right)\right)$$

Example 2.1 (Finite Experts $|F| = N$)**.** *Consider the case where ℓ is convex in its first argument and*

$$\phi(x_1, , \hat{y}_1, y_1, \ldots, x_n, \hat{y}_n, y_n) = \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) + 2\eta \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)^2 + \frac{\log N}{\eta}$$

A sufficient statistic for this problem is the $N+1$ dimensional vector given by $\mathbf{T}(x, \partial\ell(\hat{y}, y)) = (-\partial\ell(\hat{y}, y) \cdot f_1(x), \ldots, -\partial\ell(\hat{y}, y) \cdot f_N(x), \partial\ell(\hat{y}, y)^2)$ and the function V is given by

$$V(\tau) = \max_{j \in [N]} \tau[j] - 2\eta\ \tau[N+1] - \frac{\log N}{\eta}$$

To see this, notice that

$$\begin{aligned}
&\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) - 2\eta \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)^2 - \frac{\log N}{\eta} \\
&\quad = \max_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \sum_{t=1}^{n} \ell(f(x_t), y_t) - 2\eta \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)^2 - \frac{\log N}{\eta} \\
&\quad \leq \max_{f \in \mathcal{F}} \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)(\hat{y}_t - f(x_t)) - 2\eta \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)^2 - \frac{\log N}{\eta} \\
&\quad = \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + \max_{j \in [N]} \sum_{t=1}^{n} \mathbf{T}(x_t, \ell(\hat{y}_t, y_t))[j] - 2\eta \sum_{t=1}^{n} \mathbf{T}(x_t, \ell(\hat{y}_t, y_t))[N+1] - \frac{\log N}{\eta}
\end{aligned}$$

Example 2.2 (Adaptive Gradient Descent)**.** *Consider the case where ℓ is convex in its first argument and*

$$\phi(x_1, \hat{y}_1, y_1, \ldots, x_n, \hat{y}_n, y_n) = \min_{f : \|f\|_2 \leq 1} \sum_{t=1}^{n} \ell(f(x_t), y_t) + \frac{\eta}{2} \sum_{t=1}^{n} \|\nabla_t\|_2^2 + \frac{1}{2\eta}$$

A sufficient statistic for this problem is $\mathbf{T}(x_t, \partial\ell(\hat{y}_t, y_t)) = (\|\partial\ell(\hat{y}_t, y_t) \cdot x_t\|_2^2, \partial\ell(\hat{y}_t, y_t) \cdot x_t)$ and the function V is given by

$$V(\tau) = \|\tau[2 : d+1]\|_2 - \frac{\eta}{2}\tau[1] - \frac{1}{2\eta}$$

To see this, notice that

$$\begin{aligned}
&\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f:\|f\|_2\le 1} \sum_{t=1}^{n} \ell(f^\top x_t, y_t) - \frac{\eta}{2}\sum_{t=1}^{n}\|\nabla_t\|_2^2 - \frac{1}{2\eta} \\
&\qquad = \sup_{f:\|f\|_2\le 1} \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \sum_{t=1}^{n} \ell(f^\top x_t, y_t) - \frac{\eta}{2}\sum_{t=1}^{n}\|\nabla_t\|_2^2 - \frac{1}{2\eta} \\
&\qquad \le \sup_{f:\|f\|_2\le 1} \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)(\hat{y}_t - f^\top x_t) - \frac{\eta}{2}\sum_{t=1}^{n}\|\nabla_t\|_2^2 - \frac{1}{2\eta} \\
&\qquad = \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)\cdot\hat{y}_t + \sup_{f:\|f\|_2\le 1} f^\top\left(-\sum_{t=1}^{n}\partial\ell(\hat{y}_t, y_t)\cdot x_t\right) - \frac{\eta}{2}\sum_{t=1}^{n}\|\nabla_t\|_2^2 - \frac{1}{2\eta} \\
&\qquad = \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)\cdot\hat{y}_t + \left\|\sum_{t=1}^{n}\partial\ell(\hat{y}_t, y_t)\cdot x_t\right\|_2 - \frac{\eta}{2}\sum_{t=1}^{n}\|\nabla_t\|_2^2 - \frac{1}{2\eta} \\
&\qquad = \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)\cdot\hat{y}_t + \left\|\sum_{t=1}^{n}\mathbf{T}(x_t, \partial\ell(\hat{y}_t, y_t))[2:d+1]\right\|_2 - \frac{\eta}{2}\sum_{t=1}^{n}\mathbf{T}(x_t, \partial\ell(\hat{y}_t, y_t))[1] - \frac{1}{2\eta}
\end{aligned}$$

3 Burkholder Mappings

Now we define the notion of Burkholder functions that will be useful for developing algorithms.

Definition 2. *A function $\mathbf{U} : \mathcal{T} \mapsto \mathbb{R}$ along with mapping $\mathbf{T} : \mathcal{X} \times \mathbb{R} \mapsto \mathcal{T}$, is said to be a Burkholder function if it satisfies the following properties:*

1. $U(0) \le 0$
2. *For any $\tau \in \mathcal{T}$, any $x \in \mathcal{X}$ and any distribution p such that $\mathbb{E}_{\alpha\sim p}[\alpha] = 0$:*

$$\mathbb{E}_{\alpha\sim p} U(\tau + \mathbf{T}(x, \alpha)) \le U(\tau) \qquad \textit{(Restricted Concavity)}$$

The following lemma says that if a sufficient statistic turns out to be a Burkholder function, then one can design online learning algorithm using such a Burkholder function.

Lemma 1. *If $(U, \mathbf{T})$ is a sufficient statistic w.r.t. ϕ, and the function U is a Burkholder function w.r.t. mapping $\mathbf{T}$, then there exists an algorithm such that, for any n,*

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t\sim q_t}\left[\ell(\hat{y}_t, y_t)\right] \le \frac{1}{n}\phi(x_1, \hat{y}_1, y_1, \ldots, x_n, \hat{y}_n, y_n)$$

specifically, given $\tau_{t-1} = \sum_{s=1}^{t-1}\mathbf{T}\left(x_t, \partial\ell(\hat{y}_t, y_t)\right)$, the algorithm is given by

$$q_t = \operatorname*{argmin}_{q\in\Delta(\hat{Y})} \sup_{y_t\in\mathcal{Y}} \mathbb{E}_{\hat{y}_t\sim q}\left[\partial\ell(\hat{y}_t, y_t)\cdot\hat{y}_t + U\left(\tau_{t-1} + \mathbf{T}\left(x_t, \partial\ell(\hat{y}_t, y_t)\right)\right)\right] \tag{1}$$

Proof. Note that since $(U, \mathbf{T})$ is a sufficient statistic w.r.t. ϕ,

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \phi(x_1, \hat{y}_1, y_1, \ldots, x_n, \hat{y}_n, y_n) \leq \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + U\left(\sum_{t=1}^{n} \mathbf{T}\left(x_t, \partial\ell(\hat{y}_t, y_t)\right)\right)$$

Hence we can conclude that there exists a strategy that can guarantee that $\mathbb{E}\left[\sum_{t=1}^{n} \ell(\hat{y}_t, y_t)\right] \leq \phi(x_1, y_1, \ldots, x_n, y_n)$ if we can conclude that for the prescribed strategy,

$$\mathbb{E}\left[\sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + U\left(\sum_{t=1}^{n} \mathbf{T}\left(x_t, \partial\ell(\hat{y}_t, y_t)\right)\right)\right] \leq 0$$

For any t, let $\tau_t = \sum_{s=1}^{t} \mathbf{T}\left(x_s, \partial\ell(\hat{y}_s, y_s)\right)$. Note that for any n, using the prescribed strategy,

$$\begin{aligned}
&\mathbb{E}_{\hat{y}_n \sim q_n}\left\{\partial\ell(\hat{y}_n, y_n) \cdot \hat{y}_n + U\left(\sum_{t=1}^{n} \mathbf{T}\left(x_t, \partial\ell(\hat{y}_t, y_t)\right)\right)\right\} \qquad (2)\\
&= \mathbb{E}_{\hat{y}_n \sim q_n}\left\{\partial\ell(\hat{y}_n, y_n) \cdot \hat{y}_n + U\left(\tau_n\right)\right\}\\
&\leq \sup_{y_n} \mathbb{E}_{\hat{y}_n \sim q_n}\left\{\partial\ell(\hat{y}_n, y_n) \cdot \hat{y}_n + U\left(\tau_{n-1} + \mathbf{T}\left(x_n, \partial\ell(\hat{y}_n, y_n)\right)\right)\right\}\\
&= \inf_{q_n \in \Delta(\hat{Y})} \sup_{y_n} \mathbb{E}_{\hat{y}_n \sim q_n}\left\{\partial\ell(\hat{y}_n, y_n) \cdot \hat{y}_n + U\left(\tau_{n-1} + \mathbf{T}\left(x_n, \partial\ell(\hat{y}_n, y_n)\right)\right)\right\}\\
&= \inf_{q_n \in \Delta(\hat{Y})} \sup_{p_n \in \Delta(\mathcal{Y})} \mathbb{E}_{y_n \sim p_n} \mathbb{E}_{\hat{y}_n \sim q_n}\left\{\partial\ell(\hat{y}_n, y_n) \cdot \hat{y}_n + U\left(\tau_{n-1} + \mathbf{T}\left(x_n, \partial\ell(\hat{y}_n, y_n)\right)\right)\right\}\\
&= \sup_{p_n \in \Delta(\mathcal{Y})} \inf_{\hat{y}_n \in \hat{Y}} \mathbb{E}_{y_n \sim p_n}\left\{\partial\ell(\hat{y}_n, y_n) \cdot \hat{y}_n + U\left(\tau_{n-1} + \mathbf{T}\left(x_n, \partial\ell(\hat{y}_n, y_n)\right)\right)\right\}
\end{aligned}$$

Where in the above we used minimax theorem to swap min and max. Now in the last line above, lets use $\hat{y}_n^* = \operatorname*{argmin}_{\hat{y}} \mathbb{E}_{y_n \sim p_n}[\ell(\hat{y}, y_n)]$ so that $\mathbb{E}_{y_n \sim p_n}[\partial\ell(\hat{y}, y_n)] = 0$. Hence we have,

$$\begin{aligned}
&\mathbb{E}_{\hat{y}_n \sim q_n}\left\{\partial\ell(\hat{y}_n, y_n) \cdot \hat{y}_n + U\left(\sum_{t=1}^{n} \mathbf{T}\left(x_t, \partial\ell(\hat{y}_t, y_t)\right)\right)\right\}\\
&\leq \sup_{p_n \in \Delta(\mathcal{Y})} \mathbb{E}_{y_n^* \sim p_n}\left\{\partial\ell(\hat{y}_n^*, y_n) \cdot \hat{y}_n^* + U\left(\tau_{n-1} + \mathbf{T}\left(x_n, \partial\ell(\hat{y}_n, y_n)\right)\right)\right\}\\
&= \sup_{p_n \in \Delta(\mathcal{Y})} \left\{\mathbb{E}_{y_n^* \sim p_n}\left[\partial\ell(\hat{y}_n^*, y_n)\right] \cdot \hat{y}_n^* + \mathbb{E}_{y_n^* \sim p_n}\left[U\left(\tau_{n-1} + \mathbf{T}\left(x_n, \partial\ell(\hat{y}_n^*, y_n)\right)\right)\right]\right\}\\
&= \sup_{p_n \in \Delta(\mathcal{Y})} \left\{\mathbb{E}_{y_n^* \sim p_n}\left[U\left(\tau_{n-1} + \mathbf{T}\left(x_n, \partial\ell(\hat{y}_n^*, y_n)\right)\right)\right]\right\}\\
&\leq \sup_{P_n : \mathbb{E}_{\alpha \sim P_n}[\alpha] = 0} \mathbb{E}_{\alpha_n \sim P_n}\left[U\left(\tau_{n-1} + \mathbf{T}\left(x_n, \alpha_n\right)\right)\right] \leq U\left(\tau_{n-1}\right)
\end{aligned}$$

where the last line used the restricted concavity property. Thus using this recursively, we conclude:

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + U\left(\tau_n\right)\right] &= \mathbb{E}\left[\sum_{t=1}^{n-1} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + \mathbb{E}_{\hat{y}_n \sim q_n}\left[\partial\ell(\hat{y}_n, y_n) \cdot \hat{y}_n + U\left(\tau_{n-1} + \mathbf{T}\left(x_n, \partial\ell(\hat{y}_n, y_n)\right)\right)\right]\right]\\
&\leq \mathbb{E}\left[\sum_{t=1}^{n-1} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + U\left(\tau_{n-1}\right)\right] \leq \ldots\\
&\leq \mathbb{E}\left[\sum_{t=1}^{n-2} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + U\left(\tau_{n-2}\right)\right] \leq \ldots \leq U\left(0\right) \leq 0
\end{aligned}$$

Notice that this proof is similar to that of relaxations. The key though algorithmically, is that at time t, the only information about the past that the algorithm needs to solve the optimization is the vector $\tau_{t-1} = \sum_{s=1}^{t-1} \mathbf{T}(x_s, \partial\ell(\hat{y}_s, y_s))$ which on every round t can be updated by only computing $\mathbf{T}(x_t, \partial\ell(\hat{y}_t, y_t))$ and adding to the vector from the previous round. □

3.1 Fast Implementation

If the mapping $\alpha \mapsto U(\tau, \mathbf{T}(x, \alpha))$ is a convex mapping for all τ, x then it turns out that the Burkholder algorithm is simple to implement. The following proposition elucidates this.

Proposition 2. *Assume that the mapping $\alpha \mapsto U(\tau, \mathbf{T}(x, \alpha))$ is convex and that $|\partial\ell(\hat{y}, y)| \leq L$, then the following algorithm provides the bound on prediction error.*

$$\hat{y}_t = \frac{1}{2L}\left(U(\tau_{t-1}, \mathbf{T}(x, -L)) - U(\tau_{t-1}, \mathbf{T}(x, +L))\right)$$

Proof. We just need to modify the proof of lemma 1. Specifically, starting from Eq. 2, note that

$$\begin{aligned}
&\partial\ell(\hat{y}_t, y_t)\cdot\hat{y}_t + U(\tau_{t-1}, \mathbf{T}(x_t, \partial\ell(\hat{y}_t, y_t))) \\
&\qquad \leq \sup_{\partial_t \in [-L,L]} \{\partial_t \hat{y}_t + U(\tau_{t-1}, \mathbf{T}(x_t, \partial_t))\}
\end{aligned}$$

Convexity in ∂_t implies that the supremum above is achieved at either $+L$ or $-L$ and so

$$= \max_{\partial_t \in \{-L,L\}} \{\partial_t \hat{y}_t + U(\tau_{t-1}, \mathbf{T}(x_t, \partial_t))\}$$

Plugging in the strategy $\hat{y}_t$,

$$\begin{aligned}
&= \max_{\partial_t \in \{-L,L\}} \left\{ \frac{\partial_t}{2L}\left(U(\tau_{t-1}, \mathbf{T}(x, -L)) - U(\tau_{t-1}, \mathbf{T}(x, +L))\right) + U(\tau_{t-1}, \mathbf{T}(x_t, \partial_t)) \right\} \\
&= \max_{\epsilon \in \{-1,1\}} \{\epsilon\left(U(\tau_{t-1}, \mathbf{T}(x, -L)) - U(\tau_{t-1}, \mathbf{T}(x, +L))\right) + U(\tau_{t-1}, \mathbf{T}(x_t, \epsilon L))\} \\
&= \frac{U(\tau_{t-1}, \mathbf{T}(x, -L)) + U(\tau_{t-1}, \mathbf{T}(x, +L))}{2} \\
&= \mathbb{E}_\epsilon U(\tau_{t-1}, \mathbf{T}(x, \epsilon L)) \leq U(\tau_{n-1})
\end{aligned}$$

where the last inequality is via the Restricted concavity property. The rest of the proof is same as in Lemma 1. □

4 Examples

Example 2.2 Consider the sufficient statistics pair $(V, \mathbf{T})$ from Example 2.2 give by $\mathbf{T}(x_t, \partial\ell(\hat{y}_t, y_t)) = (\|\partial\ell(\hat{y}_t, y_t)\cdot x_t\|_2^2, \partial\ell(\hat{y}_t, y_t)\cdot x_t)$ and

$$V(\tau) = \|\tau[2:d+1]\|_2 - \frac{\eta}{2}\tau[1] - \frac{1}{2\eta}$$

The function $U(\tau) = \frac{\eta}{2}\|\tau[2:d+1]\|_2^2 - \frac{\eta}{2}\tau[1]$ is such that U is a Burkholder function and $U(\tau) \geq V(\tau)$ for any τ. Hence $(U, \mathbf{T})$ is also a sufficient statistic and is also a Burkholder mapping.

Proposition 3. *$U(\tau) = \frac{\eta}{2}\|\tau[2:d+1]\|_2^2 - \frac{\eta}{2}\tau[1]$ is a Burkholder function w.r.t. $\mathbf{T}$ and is a sufficient statistic for Example 2.2.*

Proof. To prove that $U, \mathbf{T}$ is a sufficient statistic, all we need to do is show that for any τ, $V(\tau) \leq U(\tau)$. To this end note that,

$$\begin{aligned} V(\tau) &= \|\tau[2:d+1]\|_2 - \frac{\eta}{2}\tau[1] - \frac{1}{2\eta} \\ &\leq \frac{\eta}{2}\|\tau[2:d+1]\|_2^2 + \frac{1}{2\eta} - \frac{\eta}{2}\tau[1] - \frac{1}{2\eta} \\ &= \frac{\eta}{2}\|\tau[2:d+1]\|_2^2 - \frac{\eta}{2}\tau[1] = U(\tau) \end{aligned}$$

That $U(0) \leq 0$ is obvious. Finally, to show that U is Burkholder, note that for any $\tau \in \mathcal{T}$, $x \in \mathcal{X}$ and distribution p s.t. $\mathbb{E}_{\alpha\sim p}[\alpha] = 0$,

$$\begin{aligned} \mathbb{E}_\alpha U(\tau + \mathbf{T}(x,\alpha)) &= \mathbb{E}_\alpha\left[\frac{\eta}{2}\|\tau[2:d+1] + \alpha x\|_2^2 - \frac{\eta}{2}\alpha^2\|x\|_2^2 - \frac{\eta}{2}\tau[1]\right] \\ &= \mathbb{E}_\alpha\left[\frac{\eta}{2}\|\tau[2:d+1]\|^2 + \frac{\eta}{2}\alpha^2\|x\|_2^2 + \eta\alpha x^\top\tau[2:d+1] - \frac{\eta}{2}\alpha^2\|x\|_2^2 - \frac{\eta}{2}\tau[1]\right] \\ &= \frac{\eta}{2}\|\tau[2:d+1]\|^2 + \eta\mathbb{E}_\alpha[\alpha]\, x^\top\tau[2:d+1] - \frac{\eta}{2}\tau[1] = U(\tau) \end{aligned}$$

□

Thus, this $U, \mathbf{T}$ function is a Burkholder mapping and is a sufficient statistic and hence can be used to compute predictions as specified by Eq. 1.

Example: 2.1 Consider the sufficient statistics pair $(V, \mathbf{T})$ from Example 2.1 given by $\mathbf{T}(x, \partial\ell(\hat{y}, y)) = (-\partial\ell(\hat{y}, y)\cdot f_1(x), \ldots, -\partial\ell(\hat{y}, y)\cdot f_N(x), \partial\ell(\hat{y}, y)^2)$ and

$$V(\tau) = \max_{j\in[N]} \tau[j] - 2\eta\ \tau[N+1] - \frac{\log N}{\eta}$$

For this example, we claim that the function $U(\tau) = \frac{1}{\eta}\log\left(\sum_{j=1}^N \exp(\eta\tau[j])\right) - 2\eta\tau[N+1] - \frac{\log N}{\eta}$ is a Burkholder function w.r.t. $\mathbf{T}$ and is such that $U(\tau) \geq V(\tau)$ for any τ. Hence $(U, \mathbf{T})$ is also a sufficient statistic and is also a Burkholder mapping.

Proposition 4. *The mapping $U(\tau) = \frac{1}{\eta}\log\left(\sum_{j=1}^N \exp(\eta\tau[j])\right) - 2\eta\tau[N+1] - \frac{\log N}{\eta}$ is a Burkholder function w.r.t. $\mathbf{T}$ and is a sufficient statistic for Example 2.1.*

Proof. To show that $U(\tau) \geq V(\tau)$, note that,

$$\begin{aligned} V(\tau) &= \max_{j\in[N]} \tau[j] - 2\eta\ \tau[N+1] - \frac{\log N}{\eta} \\ &\leq \frac{1}{\eta}\log\left(\sum_{j=1}^N \exp(\eta\tau[j])\right) - 2\eta\tau[N+1] - \frac{\log N}{\eta} = U(\tau) \end{aligned}$$

Next, note that,

$$U(0) = \frac{1}{\eta}\log\left(\sum_{j=1}^{N}\exp(0)\right) - \frac{\log N}{\eta} = \frac{1}{\eta}\log\left(N\right) - \frac{\log N}{\eta} = 0$$

Now to conclude that U is a Burkholder mapping, we need to show that $\mathbb{E}_{\alpha}[U(\tau + T(x,\alpha)] \leq U(\tau)$ for any 0 mean R.V. α. To this end, note that for any 0 mean distribution p,

$$\begin{aligned}
&\mathbb{E}_{\alpha\sim p}[U(\tau + T(x,\alpha)] \\
&\quad= \frac{1}{\eta}\mathbb{E}_{\alpha\sim p}\log\left(\sum_{j=1}^{N}\exp\left(\eta\tau[j] + \eta\alpha f_j(x)\right)\right) - 2\eta\tau[N+1] - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2 - \frac{\log N}{\eta} \\
&\quad= \frac{1}{\eta}\mathbb{E}_{\alpha\sim p}\log\left(\sum_{j=1}^{N}\exp\left(\eta\tau[j] + \eta\left(\alpha - \mathbb{E}_{\alpha'\sim p}[\alpha']\right) f_j(x)\right)\right) - 2\eta\tau[N+1] - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2 - \frac{\log N}{\eta}
\end{aligned}$$

in the above we use the fact that $\mathbb{E}_{\alpha'\sim p}[\alpha'] = 0$

$$\leq \frac{1}{\eta}\mathbb{E}_{\alpha,\alpha'\sim p}\log\left(\sum_{j=1}^{N}\exp\left(\eta\tau[j] + \eta\left(\alpha - \alpha'\right) f_j(x)\right)\right) - 2\eta\tau[N+1] - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2 - \frac{\log N}{\eta}$$

We can pull the expectation out using Jensen, due to convexity.

$$\begin{aligned}
&= \frac{1}{\eta}\mathbb{E}_{\alpha,\alpha'\sim p}\mathbb{E}_{\epsilon}\log\left(\sum_{j=1}^{N}\exp(\eta\tau[j] + \eta\epsilon(\alpha - \alpha')f_j(x))\right) - 2\eta - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2 - \frac{\log N}{\eta} \\
&\leq \frac{1}{\eta}\mathbb{E}_{\alpha,\alpha'\sim p}\log\left(\sum_{j=1}^{N}\mathbb{E}_{\epsilon}\exp(\eta\tau[j] + \eta\epsilon(\alpha - \alpha')f_j(x))\right) - 2\eta\tau[N+1] - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2 - \frac{\log N}{\eta} \\
&= \frac{1}{\eta}\mathbb{E}_{\alpha,\alpha'\sim p}\log\left(\sum_{j=1}^{N}\exp(\eta\tau[j])\mathbb{E}_{\epsilon}\exp(\eta\epsilon(\alpha - \alpha')f_j(x))\right) - 2\eta\tau[N+1] - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2 - \frac{\log N}{\eta}
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{\eta}\mathbb{E}_{\alpha,\alpha'\sim p}\log\left(\sum_{j=1}^{N}\exp(\eta\tau[j])\times\exp(\eta^2(\alpha-\alpha')^2 f_j^2(x)/2)\right) - 2\eta\tau[N+1] - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2 - \frac{\log N}{\eta}\\
&\leq \frac{1}{\eta}\mathbb{E}_{\alpha,\alpha'\sim p}\log\left(\sum_{j=1}^{N}\exp(\eta\tau[j])\times\exp(\eta^2(\alpha-\alpha')^2/2)\right) - 2\eta\tau[N+1] - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2\\
&\leq \frac{1}{\eta}\mathbb{E}_{\alpha,\alpha'\sim p}\log\left(\sum_{j=1}^{N}\exp(\eta\tau[j])\times\exp(\eta^2\alpha^2+\eta^2(\alpha')^2)\right) - 2\eta\tau[N+1] - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2 - \frac{\log N}{\eta}\\
&= \frac{1}{\eta}\mathbb{E}_{\alpha,\alpha'\sim p}\left[\log\left(\sum_{j=1}^{N}\exp(\eta\tau[j]\right) + \eta^2\alpha^2+\eta^2(\alpha')^2)\right] - 2\eta\tau[N+1] - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2 - \frac{\log N}{\eta}\\
&= \frac{1}{\eta}\log\left(\sum_{j=1}^{N}\exp(\eta\tau[j])\right) + \eta\mathbb{E}_{\alpha\sim p}\alpha^2 + \eta\mathbb{E}_{\alpha'\sim p}(\alpha')^2 - 2\eta\tau[N+1] - 2\eta\mathbb{E}_{\alpha\sim p}\alpha^2 - \frac{\log N}{\eta}\\
&= \frac{1}{\eta}\log\left(\sum_{j=1}^{N}\exp(\eta\tau[j])\right) - 2\eta\tau[N+1] - \frac{\log N}{\eta} = U(\tau)
\end{aligned}
$$

□

Example: Diagonal Adagrad Consider the following function on $u:\mathbb{R}^2\mapsto\mathbb{R}$:

$$
u(\tau) = \begin{cases} |\tau[1]| - 2\sqrt{\tau[2]} & \text{if } |\tau[1]|^2 > \tau[2] \\ -\sqrt{2\tau[2]-\tau[1]^2} & \text{otherwise} \end{cases}
$$

First, one can easily verify that $|\tau[1]| - 2\sqrt{\tau[2]} \leq u(\tau)$ for any τ. Next, given any 0 mean random variable α, any τ and any $x\in\mathbb{R}$,

$$
\mathbb{E}u([\tau[1]+\alpha x, \tau[2]+x^2) \leq u(\tau)
$$

With this in mind, consider the mapping $U:\mathbb{R}^{2d}\mapsto\mathbb{R}$

$$
U(\tau) = \sum_{j=1}^{d} u([\tau[j], \tau[d+j]])
$$

Since $u([\tau[j],\tau[d+j]]) \geq |\tau[j]| - 2\sqrt{\tau[j+d]}$, this gives us,

$$
U(\tau) \geq \|\tau[1:d]\|_1 - 2\sum_{j=d+1}^{2d}\sqrt{\tau[j]}
$$

Further if every coordinate of $\tau[d+1:2d]$ are non-negative, we can write the above as

$$
U(\tau) \geq \|\tau[1:d]\|_1 - 2\left\|(\tau[d+1:2d])^{1/2}\right\|_1
$$

Now consider the case where we have sufficient statistic $\mathbf{T}(x, \partial\ell(\hat{y}, y)) = (\partial\ell(\hat{y}, y) \cdot x, x^2)$ and let $V(\tau) = \|\tau[1:d]\|_1 - 2\|(\tau[d+1:2d])^{1/2}\|_1$. In this case, clearly U is Burkholder w.r.t. $\mathbf{T}$ and $U(\tau) \geq V(\tau)$. Now we show that $(V, \mathbf{T})$ (and hence also $(U, \mathbf{T})$) is a sufficient statistic for the mapping

$$\phi(x_1, \hat{y}_1, y_1, \ldots, x_n, \hat{y}_n, y_n) = \min_{f:\|f\|_\infty \leq 1} \sum_{t=1}^n \ell(f(x_t), y_t) + 2 \left\| \left(\sum_{t=1}^n x_t^2 \right)^{1/2} \right\|_1$$

To see this, note that,

$$\begin{aligned}
&\sum_{t=1}^n \ell(\hat{y}_t, y_t) - \inf_{f:\|f\|_\infty \leq 1} \sum_{t=1}^n \ell(f^\top x_t, y_t) - 2 \left\| \left(\sum_{t=1}^n x_t^2 \right)^{1/2} \right\|_1 \\
&\quad = \sup_{f:\|f\|_\infty \leq 1} \sum_{t=1}^n \ell(\hat{y}_t, y_t) - \sum_{t=1}^n \ell(f^\top x_t, y_t) - 2 \left\| \left(\sum_{t=1}^n x_t^2 \right)^{1/2} \right\|_1 \\
&\quad \leq \sup_{f:\|f\|_\infty \leq 1} \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t)(\hat{y}_t - f^\top x_t) - 2 \left\| \left(\sum_{t=1}^n x_t^2 \right)^{1/2} \right\|_1 \\
&\quad = \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + \sup_{f:\|f\|_\infty \leq 1} f^\top \left(-\sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) \cdot x_t \right) - 2 \left\| \left(\sum_{t=1}^n x_t^2 \right)^{1/2} \right\|_1 \\
&\quad = \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + \left\| \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) \cdot x_t \right\|_1 - 2 \left\| \left(\sum_{t=1}^n x_t^2 \right)^{1/2} \right\|_1 \\
&\quad = \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + \left\| \sum_{t=1}^n \mathbf{T}(x_t, \partial\ell(\hat{y}_t, y_t))[1:d] \right\|_1 - 2 \left\| \left(\sum_{t=1}^n \mathbf{T}(x_t, \partial\ell(\hat{y}_t, y_t))[d+1:2d] \right)^{1/2} \right\|_1 \\
&\quad = \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + \|\tau_n[1:d]\|_1 - 2 \left\| (\tau_n[d+1:2d])^{1/2} \right\|_1 \\
&\quad \leq \sum_{t=1}^n \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + U(\tau)
\end{aligned}$$

Thus we see that U is indeed a Burkholder mapping for this ϕ.

Example: Automatically Adapting to Norm In statistical learning, when one is interested in the problem of say linear SVM or linear logistic regression etc. One can use cross validation to pick a data dependent regularization parameter or bound on radius of predictor or step sizes to use. How do we deal with this issue in an online problem? How do we pick radius of our comparator class $\mathcal{F}$ in hind sight? More specifically, what if we wanted a bound that for any $f \in \mathbb{R}^d$:

$$\sum_{t=1}^n \ell(\hat{y}_t, y_t) \leq \sum_{t=1}^n \ell(f^\top x_t, y_t) + \|f\|_2 \sqrt{n}$$

Notice that if we knew the exact f or in fact the radius of f we compare regret against, such a bound is immediate using online gradient descent by projecting to a ball of that radius. However, since we want this bound adaptively, it turns out that such a bound is in fact not possible. We need to pay something extra for adapting. What one can actually achieve (and this turns out to be minimax optimal) is a bound that for any $f \in \mathbb{R}^d$:

$$\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) \leq \sum_{t=1}^{n} \ell(f^\top x_t, y_t) + (\|f\|_2 + 1/\sqrt{n})\sqrt{2n \log(\sqrt{n}\|f\|_2 + 1)} + e$$

That is we need to pay an additional $\log(n\|f\|_2)$ actor. We can achieve this using the Burkholder method. Note that for the adaptive bound above, $\phi(x_1, y_1, \ldots, x_n, y_n) = \sum_{t=1}^{n} \ell(f^\top x_t, y_t) + (\|f\|_2 + 1/\sqrt{n})\sqrt{2n \log(\sqrt{n}\|f\|_2 + 1)} + e$. For this problem, we assume n in known in advance and we give a Burkholder function that depends on n.

Proposition 5. *Let* $\mathbf{T}(x_t, \partial\ell(\hat{y}_t, y_t)) = (1, \partial\ell(\hat{y}_t, y_t) \cdot x_t)$ *be the sufficient statistic and define*

$$U(\tau) = \frac{1}{\sqrt{n}} \exp\left(\frac{\|\tau[2:d+1]\|^2}{2\tau[1]} + \frac{1}{2}\sum_{s=\tau[1]+1}^{n} \frac{1}{s}\right) - \sqrt{e}$$

$(U, \mathbf{T})$ *is a sufficient statistic for* $\phi(x_1, y_1, \ldots, x_n, y_n) = \sum_{t=1}^{n} \ell(f^\top x_t, y_t) + (\|f\|_2 + 1/\sqrt{n})\sqrt{2n \log(\sqrt{n}\|f\|_2 + 1)} + \sqrt{e}$ *and* U *is a Burkholder mapping w.r.t.* $\mathbf{T}$.

Proof. First note that to evaluate $U(0)$ we need to evaluate $\frac{\|\tau[2:d+1]\|^2}{2\tau[1]}$ at 0 which we are going to take to be 0 (think of numerator as being squared the rate of the denominator). Hence we are left with

$$\begin{aligned} U(0) &= \frac{1}{\sqrt{n}} \exp\left(\frac{1}{2}\sum_{s=1}^{n} \frac{1}{s}\right) - \sqrt{e} \\ &\leq \frac{1}{\sqrt{n}} \exp\left(\frac{1}{2}\left(\log(n) + 1\right)\right) - \sqrt{e} = 0 \end{aligned}$$

Next, note that

$$\begin{aligned} &\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f \in \mathbb{R}^d} \sum_{t=1}^{n} \ell(f^\top x_t, y_t) - (\|f\|_2 + 1/\sqrt{n})\sqrt{2n \log(\sqrt{n}\|f\|_2 + 1)} - \sqrt{e} \\ &\quad= \sup_{f \in \mathbb{R}^d} \sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \sum_{t=1}^{n} \ell(f^\top x_t, y_t) - (\|f\|_2 + 1/\sqrt{n})\sqrt{2n \log(\sqrt{n}\|f\|_2 + 1)} - \sqrt{e} \\ &\quad\leq \sup_{f \in \mathbb{R}^d} \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)(\hat{y}_t - f^\top x_t) - (\|f\|_2 + 1/\sqrt{n})\sqrt{2n \log(\sqrt{n}\|f\|_2 + 1)} - \sqrt{e} \\ &\quad= \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + \sup_{f \in \mathbb{R}^d} f^\top \left(-\sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot x_t\right) - (\|f\|_2 + 1/\sqrt{n})\sqrt{2n \log(\sqrt{n}\|f\|_2 + 1)} - \sqrt{e} \\ &\quad= \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + \sup_{R \in \mathbb{R}^+} \left\{R\left\|\sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot x_t\right\| - (R + 1/\sqrt{n})\sqrt{2n \log(\sqrt{n}R + 1)}\right\} - \sqrt{e} \\ &\quad= \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot \hat{y}_t + \sup_{\alpha \in \mathbb{R}^+} \left\{\frac{\alpha}{\sqrt{n}}\left\|\sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t) \cdot x_t\right\| - (\alpha + 1)\sqrt{2\log(\alpha + 1)}\right\} - \sqrt{e} \end{aligned}$$

Solving for α is a quadrating equation. Finally massaging the terms we can arrive at upper bound:

$$\begin{aligned}
\sum_{t=1}^{n} \ell(\hat{y}_t, y_t) - \inf_{f\in\mathbb{R}^d} \sum_{t=1}^{n} \ell(f^\top x_t, y_t) - (\|f\|_2 + 1/\sqrt{n})\sqrt{2n\log(\sqrt{n}\|f\|_2 + 1)} - \sqrt{e} \\
&\leq \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)\cdot \hat{y}_t + \frac{1}{\sqrt{n}} \exp\left(\frac{\|\sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)\cdot x_t\|^2}{2n}\right) - \sqrt{e} \\
&= \sum_{t=1}^{n} \partial\ell(\hat{y}_t, y_t)\cdot \hat{y}_t + U(\sum_{t=1}^{n} \mathbf{T}(x_t, \partial\ell(\hat{y}_t, y_t)))
\end{aligned}$$

Finally note that,

$$\begin{aligned}
U(\tau + \mathbf{T}(x,\alpha)) &= \frac{1}{\sqrt{n}} \exp\left(\frac{\|\tau[2:d+1] + \alpha x\|^2}{2(\tau[1]+1)} + \frac{1}{2}\sum_{s=\tau[1]+2}^{n} \frac{1}{s}\right) - \sqrt{e} \\
&= \frac{1}{\sqrt{n}} \exp\left(\frac{\|\tau[2:d+1]\|^2 + 2\alpha x^\top \tau[2:d+1] + \alpha^2\|x\|^2}{2(\tau[1]+1)} + \frac{1}{2}\sum_{s=\tau[1]+2}^{n} \frac{1}{s}\right) - \sqrt{e}
\end{aligned}$$

Now note that $U(\tau + T(x,\alpha))$ is convex in α and hence, if $\alpha \in [-1,1]$, by using Jensen, we can always take the worst distribution over α to be ϵ where ϵ is coin flip.

$$\begin{aligned}
\mathbb{E}_\alpha U(\tau + \mathbf{T}(x,\alpha)) &\leq \frac{1}{\sqrt{n}}\mathbb{E}_\epsilon \exp\left(\frac{\|\tau[2:d+1]\|^2 + 2\epsilon x^\top \tau[2:d+1] + \|x\|^2}{2(\tau[1]+1)} + \frac{1}{2}\sum_{s=\tau[1]+2}^{n} \frac{1}{s}\right) - \sqrt{e} \\
&\leq \frac{1}{\sqrt{n}} \exp\left(\frac{\|\tau[2:d+1]\|^2 + \|x\|^2}{2(\tau[1]+1)} + \frac{(x^\top \tau[2:d+1])^2}{2(\tau[1]+1)^2} + \frac{1}{2}\sum_{s=\tau[1]+2}^{n} \frac{1}{s}\right) - \sqrt{e} \\
&\leq \frac{1}{\sqrt{n}} \exp\left(\frac{\|\tau[2:d+1]\|^2 + 1}{2(\tau[1]+1)} + \frac{\|\tau[2:d+1])\|^2}{2(\tau[1]+1)^2} + \frac{1}{2}\sum_{s=\tau[1]+2}^{n} \frac{1}{s}\right) - \sqrt{e} \\
&\leq \frac{1}{\sqrt{n}} \exp\left(\frac{\|\tau[2:d+1]\|^2}{2}\left(\frac{1}{(\tau[1]+1)} + \frac{1}{(\tau[1]+1)^2}\right) + \frac{1}{2}\sum_{s=\tau[1]+1}^{n} \frac{1}{s}\right) - \sqrt{e} \\
&\leq \frac{1}{\sqrt{n}} \exp\left(\frac{\|\tau[2:d+1]\|^2}{2\tau[1]} + \frac{1}{2}\sum_{s=\tau[1]+1}^{n} \frac{1}{s}\right) - \sqrt{e}
\end{aligned}$$

This completes the proof. □

Example: Matrix Completion

5 Doubling Trick To Get Optimal η

Assume loss of our algorithm on any round is bounded by B and assume that the loss is non-negative. Notice that in Example 2.1 and 2.2 and many others, the ϕ function we consider is or

form

$$\phi_\eta(x_1, \hat{y}_1, y_1, \ldots, x_n, \hat{y}_n, y_n) = \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) + \frac{\eta}{2}\Psi(x_1, y_1, \hat{y}_1, \ldots, x_n, y_n, \hat{y}_n) + \frac{A}{2\eta}$$

where $\Psi(x_1, y_1, \hat{y}_1, \ldots, x_n, y_n, \hat{y}_n)$ is a non-decreasing function. In an ideal world, instead of picking a particular η, one might want to bound average expected loss by the min over η above. We will see that if one has the right algorithm for every given η, then using doubling trick, once can obtain the bound associated with taking minimum over η.

The prediction problem is broken into phases, with a constant learning rate $\eta_i = \eta_0 2^{-i}$ throughout the i-th phase (we restart the algorithm at phase i with new η_i), for some $\eta_0 > 0$. Define for $i \geq 1$

$$s_{i+1} = \min\{\tau : \eta_i \Psi(x_{s_i:\tau}, y_{s_i:\tau}, \hat{y}_{s_i:\tau}) > A/\eta_i\}$$

to be the start of the phase $i+1$, and $s_1 = 0$. Let N be the last phase of the game and let $s_{N+1} = n$. Without loss of generality, assume $N > 1$ (for, otherwise ϕ_{η_1} is at most $\min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) + 2A/\eta_0$). Now note that,

$$\begin{aligned}
\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t}\left[\ell(\hat{y}_t, y_t)\right] &= \sum_{k=1}^{N} \sum_{t=s_k+1}^{s_{k+1}} \mathbb{E}_{\hat{y}_t \sim q_t}\left[\ell(\hat{y}_t, y_t)\right] \\
&= \sum_{k=1}^{N} \sum_{t=s_k+1}^{s_{k+1}-1} \mathbb{E}_{\hat{y}_t \sim q_t}\left[\ell(\hat{y}_t, y_t)\right] + \sum_{k=1}^{N} \mathbb{E}_{\hat{y}_{s_{k+1}} \sim q_{s_{k+1}}}\left[\ell(\hat{y}_{s_{k+1}}, y_{s_{k+1}})\right] \\
&\leq \sum_{k=1}^{N} \phi_{\eta_k}(x_{s_k+1:s_{k+1}-1}, \hat{y}_{s_k+1:s_{k+1}-1}, y_{s_k+1:s_{k+1}-1}) + NB \\
&= \sum_{k=1}^{N} \min_{f \in \mathcal{F}} \sum_{t=s_k+1}^{s_{k+1}-1} \ell(f(x_t), y_t) + \frac{\eta_k}{2}\Psi(x_{s_k+1:s_{k+1}-1}, \hat{y}_{s_k+1:s_{k+1}-1}, y_{s_k+1:s_{k+1}-1}) + \frac{A}{2\eta_k} + NB \\
&\leq \min_{f \in \mathcal{F}} \sum_{k=1}^{N} \sum_{t=s_k+1}^{s_{k+1}-1} \ell(f(x_t), y_t) + \sum_{k=1}^{N} \left(\frac{\eta_k}{2}\Psi(x_{s_k+1:s_{k+1}-1}, \hat{y}_{s_k+1:s_{k+1}-1}, y_{s_k+1:s_{k+1}-1}) + \frac{A}{2\eta_k} \right) + NB \\
&\leq \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) + \sum_{k=1}^{N} \frac{A}{\eta_k} + NB
\end{aligned}$$

where the last step is because within each phase, $\eta_k \Psi(x_{s_k:s_{k+1}-1}, \hat{y}_{s_k:s_{k+1}-1}, y_{s_k:s_{k+1}-1}) \leq A/\eta_k$. Hence we have that,

$$\begin{aligned}
\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t}\left[\ell(\hat{y}_t, y_t)\right] &\leq \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) + \frac{A}{\eta_0} \sum_{k=1}^{N} 2^k + NB \\
&\leq \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) + \frac{A 2^N}{\eta_0} + NB \\
&= \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) + \frac{2A}{\eta_0} 2^{N-1} + NB
\end{aligned}$$

Now note that $NB \leq B2^{N-1}$. Choose $\eta_0 = A/B$ so that $NB \leq A2^{N-1}/\eta_0$. Hence:

$$\leq \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) + \frac{3A}{\eta_0} 2^{N-1}$$

Observe that when the $N-1$-th phase switches to the N-th, we have, $\eta_{N-1}\Psi(x_{s_{N-1}:s_N}, y_{s_{N-1}:s_N}, \hat{y}_{s_{N-1}:s_N}) > A/\eta_{N-1}$ and hence, $2^{2(N-1)} \leq \frac{\eta_0^2 \Psi(x_{s_{N-1}:s_N}, y_{s_{N-1}:s_N}, \hat{y}_{s_{N-1}:s_N})}{A} \leq \frac{\eta_0^2 \Psi(x_{1:n}, y_{1:n}, \hat{y}_{1:n})}{A}$. Hence plugging this in in the above inequality we get:

$$\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t} \left[\ell(\hat{y}_t, y_t)\right] \leq \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(f(x_t), y_t) + 3\sqrt{A\ \Psi(x_{1:n}, y_{1:n}, \hat{y}_{1:n})}$$