Learning with Complex Loss Functions and Constraints

Harikrishna Narasimhan
Institute for Applied Computational Science, SEAS, Harvard University, Cambridge, MA 02138, USA

Appendix

A Proofs

A.1 Proof of Theorem 1

Theorem (Regret Bound for COCO (Restated))**.** *For any $\delta \in (0,1]$, let the following hold w.p. $\geq 1-\delta$ (over $S \sim D^m$): for each iteration $t \in [T]$, the Frank-Wolfe algorithm satisfies $\mathcal{L}(\widehat{C}^t, \lambda^t) \leq \min_{C\in\mathcal{C}_D} \mathcal{L}(C, \lambda^t) + \theta(\delta, m)$, and $\|C^D[\widehat{h}^t] - \widehat{C}^t\|_\infty \leq \xi(\delta, m)$, where $\theta, \xi : (0,1] \times \mathbb{N} \to \mathbb{R}_+$. Let parameter B be s.t. $B \geq 2\max_{k\in[K]} \lambda^*_k$. Let $\bar{h} : \mathcal{X} \to \Delta_n$ be the classifier obtained after $T = \tau m$ iterations, for some $\tau \in \mathbb{N}$. Then w.p. $\geq 1-\delta$ (over $S \sim D^m$):*

$$L(\bar{h}) \,\leq\, L(h^*) + \frac{KB^2 + 2KR^2}{2\sqrt{\tau m}} + \theta(\delta, m) + G\xi(\delta, m)$$

and $\forall k \in [K]$,

$$g_k(\bar{h}) \,\leq\, \epsilon_k + \frac{2}{B}\left(\frac{KB^2 + 2KR^2}{2\sqrt{\tau m}} + \theta(\delta, m)\right) + \, G\xi(\delta, m).$$

For ease of presentation, we will work with constraints of the form $\phi_k(C) \leq 0$, with the constant ϵ_k absorbed into ϕ_k. We will find it useful to prove the following lemma:

Lemma 5. *For any $\delta \in (0,1]$, w.p. $\geq 1-\delta$,*

$$\max_{\lambda\in[0,B]^K} \mathcal{L}(\bar{C}, \lambda) \,-\, \min_{C\in\mathcal{C}_D} \mathcal{L}(C, \bar{\lambda}) \,\leq\, \frac{KB^2 + 2KR^2}{2\sqrt{T}} \,+\, \theta(\delta, m)$$

Proof. Following standard online gradient ascent analysis to the sequence of functions $\mathcal{L}(C^1, \lambda), \ldots, \mathcal{L}(\widehat{C}^t, \lambda)$ linear in λ, we get after T iterations:

$$\max_{\lambda\in[0,B]^K} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\widehat{C}^t, \lambda) \,-\, \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\widehat{C}^t, \lambda^t) \quad\leq\quad \frac{KB^2 + 2KR^2}{2\sqrt{T}} \tag{3}$$

where we use the fact that $\|\lambda\|_2^2 \leq KB^2$ and $\mathcal{L}$ is Lipschtiz in λ w.r.t. ℓ_2 norm with parameter $\sqrt{K}R$. We then have

$$\begin{aligned}
\max_{\lambda\in[0,B]^K} \mathcal{L}(\bar{C}, \lambda) \,-\, \min_{C\in\mathcal{C}_D} \mathcal{L}(C, \bar{\lambda}) &= \max_{\lambda\in[0,B]^K} \mathcal{L}(\bar{C}, \lambda) \,-\, \min_{C\in\mathcal{C}_D} \mathcal{L}(C, \bar{\lambda}) \\
&\leq \max_{\lambda\in[0,B]^K} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\widehat{C}^t, \lambda) \,-\, \min_{C\in\mathcal{C}_D} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(C, \lambda^t) \\
&\leq \max_{\lambda\in[0,B]^K} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\widehat{C}^t, \lambda) \,-\, \frac{1}{T}\sum_{t=1}^{T} \min_{C\in\mathcal{C}_D} \mathcal{L}(C, \lambda^t) \\
&\leq \max_{\lambda\in[0,B]^K} \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\widehat{C}^t, \lambda) \,-\, \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}(\widehat{C}^t, \lambda^t) \,+\, \theta(\delta, m) \\
&\leq \frac{KB^2 + 2KR^2}{2\sqrt{T}} \,+\, \theta(\delta, m),
\end{aligned}$$

where the last two statement holds w.p. $\geq 1-\delta$. Here the first step follows from $\mathcal{L}$ being linear in λ and being convex in C. The fourth step follows from the Frank-Wolfe guarantee. The last step follows from (3). ☐

We are now ready to prove Theorem 1.

Proof of Theorem 1. Let (C^*, λ^*) denote an optimal solution to (OP3). Recall that C^* satisfies the constraints of the primal problem, i.e. $g(C^*) \leq 0$, and by our assumption about B, $\lambda^* \in [0, B]^K$. We get immediately from Lemma 5 w.p. $\geq 1 - \delta$,

$$\begin{aligned}
\mathcal{L}(C^*, \lambda^*) &= \max_{\lambda \in [0,B]^K} \min_{C \in \mathcal{C}_D} \mathcal{L}(C, \lambda) \geq \min_{C \in \mathcal{C}_D} \mathcal{L}(C, \bar{\lambda}) \\
&\geq \max_{\lambda \in [0,B]^K} \mathcal{L}(\bar{C}, \lambda) - \frac{KB^2 + 2KR^2}{2\sqrt{T}} - \theta(\delta, m) \\
&\geq \mathcal{L}(\bar{C}, \lambda') - \frac{KB^2 + 2KR^2}{2\sqrt{T}} - \theta(\delta, m), \qquad (4)
\end{aligned}$$

where the inequality in the last line holds for any value of $\lambda' \in [0, B]^K$.

Setting $\lambda' = \mathbf{0}$ in (4), we have w.p. $\geq 1 - \delta$

$$\begin{aligned}
\psi(\bar{C}) &\leq \psi(C^*) + \sum_{k=1}^{K} \lambda_k^* \phi_k(C^*) + \frac{KB^2 + 2KR^2}{2\sqrt{T}} + \theta(\delta, m) \\
&\leq \psi(C^*) + \frac{KB^2 + 2KR^2}{2\sqrt{T}} + \theta(\delta, m), \qquad (5)
\end{aligned}$$

where the last inequality uses the fact that $\phi_k(C^*) \leq 0$.

Since $C^D[\bar{h}] = \frac{1}{T}\sum_{t=1}^{T} C^D[h^t]$, we have:

$$\|C^D[\bar{h}] - \bar{C}\|_1 \leq \frac{1}{T}\|C^D[h^t] - \widehat{C}^t\|_1 \leq \xi(\delta, m) \qquad (6)$$

where the last inequality holds w.p. at least $1 - \delta$ for all $t \in [T]$.

It follows from (5) and (6),

$$\begin{aligned}
L(\bar{h}) &= \psi(C^D[\bar{h}]) \leq \psi(\bar{C}) + G\xi(\delta, m) \\
&\leq \psi(C^*) + \frac{KB^2 + 2KR^2}{2\sqrt{T}} + \theta(\delta, m) + G\xi(\delta, m) \\
&= L(h^*) + \frac{KB^2 + 2KR^2}{2\sqrt{T}} + \theta(\delta, m) + G\xi(\delta, m)
\end{aligned}$$

For a given $k \in [K]$, setting $\lambda'_k = \lambda^*_k + B/2$ and $\lambda'_j = \lambda^*_j$ for each $j \neq k$ in (4) (note $\lambda' \in [0, B]^K$), we have w.p. $\geq 1 - \delta$

$$\begin{aligned}
\mathcal{L}(C^*, \lambda^*) &\geq \psi(\bar{C}) + \sum_{k=1}^{K} \lambda_k^* \phi_k(\bar{C}) + \frac{B}{2}\phi_k(\bar{C}) - \frac{KB^2 + 2KR^2}{2\sqrt{T}} - \theta(\delta, m) \\
&\geq \min_{C \in \mathcal{C}_D} \left\{ \psi(C) + \sum_{k=1}^{K} \lambda_k^* \phi_k(C) \right\} + \frac{B}{2}\phi_k(\bar{C}) - \frac{KB^2 + 2KR^2}{2\sqrt{T}} - \theta(\delta, m) \\
&= \mathcal{L}(C^*, \lambda^*) + \frac{B}{2}\phi_k(\bar{C}) - \frac{KB^2 + 2KR^2}{2\sqrt{T}} - \theta(\delta, m).
\end{aligned}$$

This gives us that for each $k \in [K]$

$$\phi_k(\bar{C}) \leq \frac{2}{B}\left(\frac{KB^2 + 2KR^2}{2\sqrt{T}} + \theta(\delta, m) \right). \qquad (7)$$

It follows from (7) and (6), $\forall k \in [K]$:

$$\begin{aligned}
g_k(\bar{h}) &\leq \phi_k(C^D[\bar{h}]) \leq \phi_k(\bar{C}) + G\xi(\delta, m) \\
&\leq \frac{2}{B}\left(\frac{KB^2 + 2KR^2}{2\sqrt{T}} + \theta(\delta, m) \right) + G\xi(\delta, m).
\end{aligned}$$

Setting $T = \tau m$ completes the proof. ∎

A.2 Proof of Theorem 3

Theorem (Regret Bound for FRACO (Restated)). *Let $f'(C) \geq b$, $\forall C \in \mathcal{C}_D$ for $b > 0$. For any $\delta \in (0,1]$, w.p. $\geq 1-\delta$, in each iteration $t \in [T]$, the COCO step satisfies: $f(\widehat{C}^t) - \gamma^t f'(\widehat{C}^t) \,\leq\, \min_{C \in \mathcal{C}_D} f(C) - \gamma^t f'(C) \,+\, \theta(\delta, m)$, with each $\phi_k(\widehat{C}^t) \leq \epsilon_k + \theta'(\delta, m)$, and $\|C^D[h^t] - \widehat{C}^t\|_\infty \leq \xi(\delta, m)$. Let $\bar{h}$ be the classifier returned after $T = \tau m$ iterations. Then for any $\delta \in (0,1]$, w.p. $\geq 1-\delta$ (over $S \sim D^m$),*

$$L(\bar{h}) \;\leq\; L(h^*) \,+\, 2\,\theta(\delta,m)/b \,+\, 2G\,\xi(\delta,m)/b \,+\, 2^{-\tau m}R \quad \textit{and} \quad g_k(\bar{h}) \,\leq\, \epsilon_k \,+\, \theta'(\delta,m),\; \forall k \in [K].$$

We will find the following lemma useful.

Lemma 6. *At each iteration $t \in [T]$ of the FRACO, w.p. $\geq 1-\delta$:*

$$L(h^*) \;\geq\; \alpha^t \,-\, \frac{1}{b}\,\theta(\delta,m) \quad \textit{and} \quad L(h^t) \;\leq\; \beta^t + \frac{1}{b}\,\theta(\delta,m) + \frac{2G}{b}\,\xi(\delta,m)$$

Proof. We will use mathematical induction on the iteration number t. For $t = 0$, the invariant holds trivially: $L(h^*) \geq \alpha_0 = 0$ and $L(h^0) \leq \beta_0 = R$. Let us assume that the invariant holds at iteration $t-1$. We shall show that the invariant holds at iteration t.

For ease of presentation, henceforth, we will not explicitly qualify statements as holding with high probability. We consider two cases in line 6 of FRACO: (a) $\psi(\widehat{C}^t) \leq \gamma^t$ and (b) $\psi(\widehat{C}^t) > \gamma^t$.

Case (a): Here, $\psi(\widehat{C}^t) \leq \gamma^t$, leading to $\alpha^t = \alpha^{t-1}$, $\beta^t = \gamma^t$ and $h^t = \widehat{h}$. From our assumption that the invariant holds in iteration $t-1$, we have

$$L(h^*) \;\geq\; \alpha^{t-1} \,-\, \frac{1}{b}\,\theta(\delta,m) \;=\; \alpha^t \,-\, \frac{1}{b}\,\theta(\delta,m).$$

We also have:

$$\begin{aligned}
f(C^t) \,-\, \gamma^t f'(C^t) \;&\leq\; f(\widehat{C}^t) \,-\, \gamma^t f'(\widehat{C}^t) \,+\, 2G\,\xi(\delta,m) \\
&\leq\; \min_{C \in \mathcal{C}_D} \big\{ f(C) \,-\, \gamma^t f'(C) \big\} \,+\, \theta(\delta,m) \,+\, 2G\,\xi(\delta,m) \\
&\leq\; f(\widehat{C}^t) \,-\, \gamma^t f'(\widehat{C}^t) \,+\, \theta(\delta,m) \,+\, 2G\,\xi(\delta,m) \\
&=\; f'(\widehat{C}^t)(\psi(\widehat{C}^t) \,-\, \gamma^t) \,+\, \theta(\delta,m) \,+\, 2G\,\xi(\delta,m) \\
&\leq\; 0 \,+\, \theta(\delta,m) \,+\, 2G\,\xi(\delta,m),
\end{aligned}$$

where the first two steps uses the guarantee on COCO.

The above inequality then gives us:

$$\begin{aligned}
\frac{f(C^t)}{f'(C^t)} \;&\leq\; \gamma^t \,+\, \frac{\theta(\delta,m)}{f'(C^t)} \,+\, \frac{2G}{f'(C^t)}\,\xi(\delta,m) \\
&\leq\; \gamma^t \,+\, \frac{1}{b}\,\theta(\delta,m) \,+\, \frac{2G}{b}\,\xi(\delta,m),
\end{aligned}$$

which follows from $f'(C^t) \geq b$. Thus $\psi(C^D[h^t]) \;\leq\; \gamma^t + \frac{1}{b}\,\theta(\delta,m) + \frac{2G}{b}\,\xi(\delta,m) \;=\; \beta^t + \frac{1}{b}\,\theta(\delta,m) + \frac{2G}{b}\,\xi(\delta,m)$.

Case (b): Here $\psi(\widehat{C}^t) > \gamma^t$, leading to $\alpha^t = \gamma^t$, $\beta^t = \beta^{t-1}$ and $h^t = h^{t-1}$. We then have from the guarantee on COCO:

$$\begin{aligned}
\min_{C \in \mathcal{C}_D} f(C) \,-\, \gamma^t f'(C) \;&>\; f(\widehat{C}^t) \,-\, \gamma^t f'(\widehat{C}^t) \,-\, \theta(\delta,m) \\
&\geq\; f'(\widehat{C}^t)(\psi(\widehat{C}^t) \,-\, \gamma^t) \,-\, \theta(\delta,m) \\
&\geq\; 0 \,-\, \theta(\delta,m).
\end{aligned}$$

The above inequality then gives us for all $C \in \mathcal{C}_D$,

$$\frac{f(C)}{f'(C)} \;\geq\; \gamma^t \,-\, \frac{\theta(\delta,m)}{f'(C)}$$

$$\geq \gamma^t - \frac{1}{b}\,\theta(\delta, m).$$

Thus $\min_{C \in \mathcal{C}_D} \psi(C) > \gamma^t - \frac{1}{b}\,\theta(\delta, m) = \alpha^t - \frac{1}{b}\,\theta(\delta, m)$.

Further, by our assumption that the invariant holds at iteration $t-1$, we have

$$L(h^t) \leq \beta^{t-1} + \frac{1}{b}\,\theta(\delta, m) + \frac{2G}{b}\,\xi(\delta, m) = \beta^t + \frac{1}{b}\,\theta(\delta, m) + \frac{2G}{b}\,\xi(\delta, m).$$

□

We are now ready to prove the theorem.

Proof of Theorem 3. It is easy to show that at each iteration t:

$$\beta^t - \alpha^t = \frac{1}{2}(\beta^{t-1} - \alpha^{t-1}) \tag{8}$$

Then from Lemma 6 we have,

$$\begin{aligned} L(\bar{h}) - L(h^*) &= L(h^T) - L(h^*) \\ &\leq \beta^T - \alpha^T + \frac{2}{b}\,\theta(\delta, m) + \frac{2G}{b}\,\xi(\delta, m) \\ &\leq 2^{-T}(\beta^0 - \alpha^0) + \frac{2}{b}\,\theta(\delta, m) + \frac{2G}{b}\,\xi(\delta, m) \\ &= 2^{-T}R + \frac{2}{b}\,\theta(\delta, m) + \frac{2G}{b}\,\xi(\delta, m). \end{aligned}$$

From the guarantee for the `COCO` method, we have that $g_k(\bar{h}) \leq \theta'(\delta, m),\ \forall k \in [K]$. □

A.3 Regret Bound for Frank-Wolfe Algorithm under Fairness Constraints

We outline the variant of the `COCO` and `FrankWolfe` algorithm for a setting with fairness constraints in Algorithm 4 and 5. Here for any $u \in [M]$, we use $\text{conf}_u(h, S) \in [0,1]^{n \times n}$ to denote the *empirical* confusion matrix for a classifier h conditioned on $A = u$, from sample S:

$$[\text{conf}_u(h, S)]_{ij} = \frac{\sum_{k=1}^m \mathbf{1}(y_k = i, h(x_k) = j, a_k = u)}{\sum_{k=1}^m \mathbf{1}(a_k = u)}.$$

The following regret bound holds for the fair variant of the Frank-Wolfe algorithm.

Theorem 7 (Regret Bound for `FairFrankWolfe`)**.** *Let $\psi, \phi_1, \ldots, \phi_K([0,1]^{n\times n})^M \to \mathbb{R}_+$ be G-Lipschitz and β-smooth in $(C^1, \ldots, C^M)$ w.r.t. the ℓ_1. Let $\widehat{\eta} : \mathcal{X} \times [M] \to \Delta_n$ be the conditional class probability model used to construct the plug-in classifier for the cost-sensitive learner in line 6 of the `FairFrankWolfe`. Given $\lambda \in [0, B]^K$, let $(\widehat{C}^1, \ldots, \widehat{C}^M, \widehat{h})$ be returned by the algorithm after κm iterations for some $Q = \kappa \in \mathbb{N}$. Let $C^a = C^{D_a}[\widehat{h}]$. Then for any $\delta \in (0,1]$, w.p. $\geq 1 - \delta$ (over $S \sim D^m$)*

$$\begin{aligned} &\mathcal{L}(\widehat{C}^1, \ldots, \widehat{C}^M, \lambda) - \min_{(C^1, \ldots, C^M) \in \mathcal{C}_D} \mathcal{L}(C^1, \ldots, C^M, \lambda) \\ &\quad\leq \frac{4G(1+KB)}{\pi_{\min}} \mathbf{E}_{X,A}\big[\big\|\widehat{\eta}(X, A) - \eta(X, A)\big\|_1\big] + 4\sqrt{2}\beta(1+KB)n^2 \sum_{a=1}^M \|C^a - \widehat{C}\|_\infty + \frac{8\beta(1+KB)}{\kappa m + 2}, \end{aligned}$$

and $\forall a \in [M]$,

$$\|C^a - \widehat{C}^a\|_\infty \leq \nu \sqrt{\frac{n^2 \log(n) \log(m) + \log(n^2 M/\delta)}{m}},$$

where $\pi_{\min} = \min_{a \in [M]} \pi_a$ and $\nu > 0$ is a distribution-independent constant.

Algorithm 4 `COCO-fair`: Algorithm for Convex Losses with Convex `Fair`ness Constraints

Input: $\psi, \phi_1, \ldots, \phi_K : ([0,1]^{n\times n})^M \to \mathbb{R}_+$
$S = ((x_1, y_1, a_1), \ldots, (x_m, y_m, a_m))$
Initialize: $\lambda^0 = 0^K$, $\eta_0 > 0$
For $t = 1$ **to** $T = \tau m$ **do**
    $(\widehat{C}^{1,t}, \ldots, \widehat{C}^{M,t}, \widehat{h}^t) \leftarrow \texttt{FairFrankWolfe}(\psi, \phi_1, \ldots, \phi_K, \lambda^{t-1}, S)$
    $\lambda_k^t = \Pi_{[0,B]}\left(\lambda_k^{t-1} + \frac{\eta_0}{\sqrt{t}}\left(\phi_k(\widehat{C}^{1,t}, \ldots, \widehat{C}^{M,t}) - \epsilon_k\right)\right), \forall k$
End For
Output: Classifier $\bar{h} : \mathcal{X} \times [M] \to \Delta_n$ that for any $x \in \mathcal{X}$ and $a \in [M]$ outputs $\widehat{h}^t(x, a)$ with probability $\frac{1}{T}$

Algorithm 5 `FairFrankWolfe`: Algorithm for convex objective for the setting with fairness constraints

Input: $\psi, \phi_1, \ldots, \phi_K : ([0,1]^{n\times n})^M \to \mathbb{R}_+$, $\lambda \in \mathbb{R}_+^n$
$S = ((x_1, y_1, a_1), \ldots, (x_m, y_m, a_m))$
Split S into S_1 and S_2 with sizes $\lceil \frac{m}{2} \rceil$ and $\lfloor \frac{m}{2} \rfloor$
$\Gamma^{a,0} = \text{conf}_a(H^0, S_1), \forall a \in [M]$ for some $H^0 : \mathcal{X} \to \Delta_n$
For $r = 1$ **to** Q **do**
    $W^a = \nabla_{C^a} \psi(\Gamma^{1,r-1}, \ldots, \Gamma^{M,r-1}) + \sum_{k=1}^K \lambda_k \nabla_{C^a} \phi_k(\Gamma^{1,r-1}, \ldots, \Gamma^{M,r-1}), \;\; \forall a \in [M]$
    $H^r = \texttt{cost-sensitive}(W^1, \ldots, W^M, S_2)$
    $\Gamma^{a,r} = \left(1 - \frac{2}{r+1}\right)\Gamma^{a,r-1} + \frac{2}{r+1}\text{conf}_a(H^r, S_1), \;\; \forall a \in [M]$
End For
Output: $\widehat{C}^1 = \Gamma^{1,R}, \ldots, \widehat{C}^M = \Gamma^{M,R}$, Classifier $\widehat{h} : \mathcal{X} \times [M] \to \Delta_n$ that for $x \in \mathcal{X}$ and $a \in [M]$ outputs $H^r(x, a)$ with probability $\frac{2}{r+1} \prod_{s=r+1}^R \left(1 - \frac{2}{s+1}\right)$

It is clear from the above theorem that whens sample size $m \to \infty$, `FairFrankWolfe` method converges to the optimal objective value, provided $\mathbf{E}_X[\|\widehat{\boldsymbol{\eta}}(X) - \boldsymbol{\eta}(X)\|_1] \to 0$ as $m \to \infty$. The proof of the theorem follows the same progression as Theorem 16 in [26], except for the following lemmas.

Lemma 8 (Uniform convergence of confusion matrices)**.** *Let $\eta : \mathcal{X} \times [M] \to \Delta_n$ and let $\mathcal{H}_\eta$ be the set of (deterministic) classifiers $h : \mathcal{X} \times [M] \to [n]$ that satisfy $h(x, a) = \text{argmin}_{j \in [n]} \sum_{i=1}^n \eta_i(x, a) L_{ij}^a$ for some $\mathbf{L}^1, \ldots, \mathbf{L}^M \in [0,1]^{n\times n}$. For any $\delta \in (0, 1]$, w.p. $\leq 1 - \delta$ (over draw of $S \sim D^m$), $\forall a \in [M]$,*

$$\sup_{h \in \mathcal{H}_\eta} \|C^{D_a}[h] - \text{conf}_a(h, S)\|_\infty \leq \nu \sqrt{\frac{n^2 \log(n) \log(m) + \log(n^2 M/\delta)}{m}},$$

where $\nu > 0$ is a distribution-independent constant.

The proof of the above lemma follows by applying the uniform convergence result in [26] (see Lemma 15) for each $a \in [M]$, and taking a union bound over the M events. The next lemma bounds the regret of a plug-in classifier.

Lemma 9 (Regret of plug-in classifier)**.** *For fixed $\mathbf{L}^1, \ldots, \mathbf{L}^M \in [0,1]^{n\times n}$ define loss function $L[h] = \sum_{a=1}^M \langle \mathbf{L}^a, C^{D_a}[\widehat{h}] \rangle$. Then the following classifier is optimal for L:*

$$h^*(x, a) = \text{argmin}^*_{j \in [n]} \sum_{i=1}^n \eta_i(x, a) L_{ij}^a.$$

Moreover, given a class probability estimation model $\widehat{\eta} : \mathcal{X} \times [M] \to \Delta_n$, define a classifier:

$$\widehat{h}(x, a) = \text{argmin}_{j \in [n]} \sum_{i=1}^n \widehat{\eta}_i(x, a) L_{ij}^a.$$

Then the following is a bound on the regret of $\widehat{h}$:

$$L(\widehat{h}) - L(h^*) \leq \frac{1}{\pi_{\min}} \mathbf{E}_{X,A}\left[\|\widehat{\eta}(X, A) - \eta(X, A)\|_1\right],$$

where $\pi_{\min} = \min_{a \in [M]} \pi_a$.

Proof. Let $\boldsymbol{\ell}^a_j = \big[L^a_{1,j}, \ldots, L^a_{n,j}\big]$. We first show that h^* optimizes L:

$$\begin{aligned}\langle \mathbf{L}^a, C^{D_a}[h]\rangle &= \sum_{j=1}^n \mathbf{P}\big[Y = i, h(X,a) = j \,\big|\, A = a\big] L^a_{ij} \\ &= \mathbf{E}\bigg[\sum_{j=1}^n \eta_i(X,a)\, L^a_{i,h(X,a)} \,\bigg|\, A = a\bigg] \\ &= \mathbf{E}\bigg[\eta(X,a)^\top \boldsymbol{\ell}^a_{h(X,a)} \,\bigg|\, A = a\bigg].\end{aligned}$$

We then have

$$\begin{aligned}L(h^*) &= \sum_{a=1}^M \mathbf{E}\bigg[\eta(X,a)^\top \boldsymbol{\ell}^a_{h^*(X,a)} \,\bigg|\, A = a\bigg] \\ &= \sum_{a=1}^M \mathbf{E}\bigg[\min_{j\in[n]}\, \eta(X,a)^\top \boldsymbol{\ell}^a_j \,\bigg|\, A = a\bigg] \\ &\leq \sum_{a=1}^M \mathbf{E}\bigg[\eta(X,a)^\top \boldsymbol{\ell}^a_{h(X,a)} \,\bigg|\, A = a\bigg] \,=\, L(h),\end{aligned}$$

where the last statement holds for any classifier $h : \mathcal{X} \times [M] \to \Delta_n$. Thus $h^* \in \operatorname{argmin}_{h:\mathcal{X}\to\Delta_n} L(h)$.

We next prove the regret bound for $\widehat{h}$:

$$\begin{aligned}L(\widehat{h}) \,-\, L(h^*) &= \sum_{a=1}^M \mathbf{E}\big[\eta(X,a)^\top \boldsymbol{\ell}^a_{\widehat{h}(X,a)} \,\big|\, A = a\big] - \sum_{a=1}^M \mathbf{E}\big[\eta(X,a)^\top \boldsymbol{\ell}^a_{h^*(X,a)} \,\big|\, A = a\big] \\ &= \sum_{a=1}^M \mathbf{E}\big[\widehat{\eta}(X,a)^\top \boldsymbol{\ell}^a_{\widehat{h}(X,a)} \,+\, (\eta(X,a) - \widehat{\eta}(X,a)^\top \boldsymbol{\ell}^a_{\widehat{h}(X,a)} \,+\, \eta(X,a)^\top \boldsymbol{\ell}^a_{h^*(X,a)} \,\big|\, A = a\big] \\ &\leq \sum_{a=1}^M \mathbf{E}\big[\widehat{\eta}(X,a)^\top \boldsymbol{\ell}^a_{h^*(X,a)} \,+\, (\eta(X,a) - \widehat{\eta}(X,a))^\top \boldsymbol{\ell}^a_{\widehat{h}(X,a)} \,-\, \eta(X,a)^\top \boldsymbol{\ell}^a_{h^*(X,a)} \,\big|\, A = a\big] \\ &= \sum_{a=1}^M \mathbf{E}\big[(\eta(X,a) - \widehat{\eta}(X,a))^\top (\boldsymbol{\ell}^a_{\widehat{h}(X,a)} - \boldsymbol{\ell}^a_{h^*(X,a)}) \,\big|\, A = a\big] \\ &\leq \sum_{a=1}^M \mathbf{E}\big[\big\|\eta(X,a) - \widehat{\eta}(X,a)\big\|_1 \cdot \big\|\boldsymbol{\ell}^a_{\widehat{h}(X,a)} - \boldsymbol{\ell}^a_{h^*(X,a)}\big\|_\infty \,\big|\, A = a\big] \\ &\leq \sum_{a=1}^M \mathbf{E}\big[\big\|\eta(X,a) - \widehat{\eta}(X,a)\big\|_1 \,\big|\, A = a\big] \\ &= \sum_{a=1}^M \frac{\pi_a}{\pi_a} \mathbf{E}\big[\big\|\eta(X,a) - \widehat{\eta}(X,a)\big\|_1 \,\big|\, A = a\big] \\ &\leq \frac{1}{\pi_{\min}} \mathbf{E}_{X,A}\big[\|\eta(X,A) - \widehat{\eta}(X,A)\|_1\big],\end{aligned}$$

where the third step follows from the definition of $\widehat{h}$ and the sixth step uses the fact that $L^a_{ij} \in [0,1]$. □

The proof of Theorem 7 then follows from Lemma 9, Lemma 8, and standard convergence result for the Frank-Wolfe optimization solver for optimizing a convex objective [13]. The proof uses the fact that $\mathcal{L}(C^1, \ldots, C^M, \lambda)$ is Lipschitz w.r.t. the ℓ_1 norm with parameter $G(1 + KB)$ and smooth w.r.t. the ℓ_1 norm with parameter $\beta(1 + KB)$.