Appendices

Notations: Δ_m denotes the $(m-1)$-dimensional simplex. $[m] = \{1, 2, \ldots, m\}$ represents an index set of size m. $\|\cdot\|$ denotes the operator norm for matrices and the ℓ_2 norm for vectors. For a matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$, $\text{diag}(\mathbf{A}) = [C_{11}, \ldots, C_{mm}]^\top$ outputs the m diagonal entries. For an index $j \in [m]$, $\text{onehot}(j) \in \{0,1\}^m$ denotes a one-hot encoding of j, and for a classifier $h : \mathcal{X}\to[m]$, $\widetilde{h} = \text{onehot}(h)$ denotes the same classifier with one-hot outputs, i.e. $\widetilde{h}(x) = \text{onehot}(h(x))$.

A. Extension to General Linear Metrics

We describe how our proposal extends to black-box metrics $\mathcal{E}^D[h] = \psi(\mathbf{C}[h])$ defined by a function $\psi : [0,1]^{m\times m}\to\mathbb{R}_+$ of *all* confusion matrix entries. This handles, for example, the label noise models in Table 1 with a general (non-diagonal) noise transition matrix $\mathbf{T}$. We begin with metrics that are linear functions of the diagonal and off-diagonal confusion matrix entries $\mathcal{E}^D[h] = \sum_{ij} \beta_{ij} C_{ij}[h]$ for some $\boldsymbol{\beta} \in \mathbb{R}^{m\times m}$. In this case, we will use an example weighting function $\mathbf{W} : \mathcal{X}\to\mathbb{R}_+^{m\times m}$ that maps an instance x to an $m \times m$ weight matrix $\mathbf{W}(x)$, where $W_{ij}(x) \in \mathbb{R}_+^{m\times m}$ is the weight associated with the (i,j)-th confusion matrix entry.

Note that in practice, the metric $\mathcal{E}^D$ may depend on only a subset of d entries of the confusion matrix, in which case, the weighting function only needs to weight those entries. Consequently, the weighting function can be parameterized with Ld parameters, which can then be estimated by solving a system of Ld linear equations. For the sake of completeness, here we describe our approach for metrics that depend on all m^2 confusion entries.

Modeling weighting function: Like in (5), we propose modeling this function as a weighted sum of L basis functions:

$$W_{ij}(x) = \sum_{\ell=1}^{L} \alpha_{ij}^{\ell} \phi^{\ell}(x),$$

where each $\phi^\ell : \mathcal{X}\to[0,1]$ and $\alpha_{ij}^\ell \in \mathbb{R}$. Similar to (4), our goal is to then estimate coefficients $\boldsymbol{\alpha}$ so that:

$$\mathbf{E}_{(x,y)\sim\mu}\Big[\sum_{ij} W_{ij}(x)\,\mathbf{1}(y=i) h_j(x)\Big] \approx \mathcal{E}^D[h], \forall h. \tag{13}$$

Expanding the weighting function in (13), we get:

$$\sum_{\ell=1}^{L}\sum_{i,j} \alpha_{ij}^{\ell} \underbrace{\mathbf{E}_{(x,y)\sim\mu}\big[\phi^\ell(x)\,\mathbf{1}(y=i) h_j(x)\big]}_{\Phi_{i,j}^{\mu,\ell}[h]} \approx \mathcal{E}^D[h], \forall h,$$

which can be re-written as:

$$\sum_{\ell=1}^{L}\sum_{i,j} \alpha_{ij}^{\ell} \Phi_{ij}^{\mu,\ell}[h] \approx \mathcal{E}^D[h], \forall h. \tag{14}$$

Estimating coefficients $\boldsymbol{\alpha}$: To estimate $\boldsymbol{\alpha} \in \mathbb{R}^{Lm^2}$, our proposal is to probe the metric $\mathcal{E}^D$ at Lm^2 different classifiers $h^{\ell,1,1}, \ldots, h^{\ell,m,m}$, with one classifier for each combination (ℓ, i, j) of basis functions and confusion matrix entries, and to solve the following system of Lm^2 linear equations:

$$\begin{aligned} \sum_{\ell,i,j} \alpha_{ij}^{\ell}\, \widehat{\Phi}_{ij}^{\text{tr},\ell}[h^{1,1,1}] &= \widehat{\mathcal{E}}^{\text{val}}[h^{1,1,1}] \\ &\vdots \\ \sum_{\ell,i,j} \alpha_{ij}^{\ell}\, \widehat{\Phi}_{ij}^{\text{tr},\ell}[h^{L,m,m}] &= \widehat{\mathcal{E}}^{\text{val}}[h^{L,m,m}] \end{aligned} \tag{15}$$

Here $\widehat{\Phi}_{ij}^{\text{tr},\ell}[h]$ is an estimate of $\Phi_{ij}^{\mu,\ell}[h]$ using training sample S^{tr} and $\widehat{\mathcal{E}}^{\text{val}}[h]$ is an estimate of $\mathcal{E}^D[h]$ using the validation sample S^{val}. Equivalently, defining $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{Lm^2\times Lm^2}$ and $\widehat{\boldsymbol{\mathcal{E}}} \in \mathbb{R}^{Lm^2}$ with each:

$$\widehat{\Sigma}_{(\ell,i,j),(\ell',i',j')} = \widehat{\Phi}_{i'j'}^{\text{tr},\ell'}[h^{\ell,i,j}]; \quad \widehat{\mathcal{E}}_{(\ell,i,j)} = \widehat{\mathcal{E}}^{\text{val}}[h^{\ell,i,j}],$$

we compute $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mathcal{E}}}$.

Choosing probing classifiers: As described in Section 3.4, we propose picking each probing classifier $h^{\ell,i,j}$ so that the (ℓ, i, j)-th diagonal entry of $\widehat{\boldsymbol{\Sigma}}$ is large and the off-diagonal entries are all small. This can be framed as the following constrained satisfaction problem:

$$\text{For } h^{\ell,i,j} \text{ pick } h \in \mathcal{H} \text{ such that:}$$

$$\widehat{\Phi}^{\text{tr},\ell}_{i,j}[h] \geq \gamma, \text{ and } \widehat{\Phi}^{\text{tr},\ell'}_{i',j'}[h] \leq \omega, \forall (\ell', i', j') \neq (\ell, i, j),$$

for some $0 < \omega < \gamma < 1$. While the more practical approach prescribed in Section 3.4 of constructing the probing classifiers from trivial classifiers that predict the same class on all or a subset of examples does not apply here (because here we need to take into account both the diagonal and off-diagonal confusion entries), the above problem can be solved using off-the-shelf tools available for rate-constrained optimization problems (Cotter et al., 2019b).

Plug-in classifier: Having estimated an example weighting function $\widehat{\mathbf{W}} : \mathcal{X} \to \mathbb{R}^{m \times m}$, we seek to maximize a weighted objective on the training distribution:

$$\max_{h} \mathbf{E}_{(x,y)\sim\mu}\left[\sum_{ij} \widehat{W}_{ij}(x)\, \mathbf{1}(y = i) h_j(x)\right],$$

for which we can construct a plug-in classifier that post-shifts a pre-trained class probability model $\widehat{\eta}^{\text{tr}} : \mathcal{X} \to \Delta_m$:

$$\widehat{h}(x) \in \operatorname{argmax}_{j \in [m]} \sum_{i=1}^{m} \widehat{W}_{ij}(x)\, \widehat{\eta}^{\text{tr}}_i(x).$$

For handling general non-linear metrics $\mathcal{E}^D[h] = \psi(\mathbf{C}[h])$ with a smooth $\psi : [0,1]^{m \times m} \to \mathbb{R}_+$, we can directly adapt the iterative plug-in procedure in Algorithm 3, which would in turn construct a plug-in classifier of the above form in each iteration (line 9). See Narasimhan et al. (2015b) for more details of the iterative Frank-Wolfe based procedure for optimizing general metrics, where the authors consider non-black-box metrics in the absence of distribution shift.

B. Proofs

B.1. Proof of Theorem 1

Theorem ((Restated) **Error bound on elicited weights).** *Let the input metric be of the form $\widehat{\mathcal{E}}^{\text{lin}}[h] = \sum_i \beta_i \widehat{C}^{\text{val}}_{ii}[h]$ for some (unknown) coefficients $\boldsymbol{\beta} \in \mathbb{R}^m_+$, $\|\boldsymbol{\beta}\| \leq 1$. Let $\mathcal{E}^D[h] = \sum_i \beta_i C^D_{ii}[h]$. Let $\gamma, \omega > 0$ be such that the constraints in (11) are feasible for hypothesis class $\bar{\mathcal{H}}$, for all ℓ, i. Suppose Algorithm 1 chooses each classifier $h^{\ell,i}$ to satisfy (11), with $\mathcal{E}^D[h^{\ell,i}] \in [c, 1], \forall \ell, i$, for some $c > 0$. Let $\bar{\alpha}$ be the associated coefficient in Assumption 1 for metric $\mathcal{E}^D$. Suppose $\gamma > 2\sqrt{2}Lm\omega$ and $n^{\text{tr}} \geq \frac{L^2 m \log(Lm|\mathcal{H}|/\delta)}{(\frac{\gamma}{2} - \sqrt{2}Lm\omega)^2}$. Fix $\delta \in (0,1)$. Then w.p. $\geq 1 - \delta$ over draws of S^{tr} and S^{val} from μ and D resp., the coefficients $\widehat{\boldsymbol{\alpha}}$ output by Algorithm 1 satisfies:*

$$\|\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\| \leq \mathcal{O}\Big(\frac{Lm}{\gamma^2}\sqrt{\frac{L\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}} + \frac{\sqrt{Lm}}{\gamma}\Big(\sqrt{\frac{L^2 m \log(Lm/\delta)}{c^2\gamma^2 n^{\text{val}}}} + \nu\Big)\Big),$$

where the term $|\mathcal{H}|$ can be replaced by a measure of capacity of the hypothesis class $\mathcal{H}$.

The solution from Algorithm 1 is given by $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mathcal{E}}}$. Let $\bar{\boldsymbol{\alpha}}$ be the "true" coefficients given in Assumption 1. Let $\boldsymbol{\Sigma} \in \mathbb{R}^{Lm \times Lm}$ denote the population version of $\widehat{\boldsymbol{\Sigma}}$, with $\Sigma_{(\ell,i),(\ell',i')} = \mathbf{E}_{(x,y)\sim\mu}\big[\phi^{\ell'}(x)\mathbf{1}(y = i')h^{\ell,i}_{i'}(x)\big]$. Similarly, denote the population version of $\widehat{\boldsymbol{\mathcal{E}}}$ by: $\mathcal{E}_{(\ell,i)} = \mathcal{E}^D[h^{\ell,i}]$. Let $\boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}}$ be the solution we obtain had we used the population versions of these quantities. Further, define the vector $\bar{\boldsymbol{\mathcal{E}}} \in \mathbb{R}^{Lm}$:

$$\bar{\mathcal{E}}_{(\ell',i')} = \sum_{\ell,i} \bar{\alpha}^{\ell}_i \Phi^{\mu,\ell}_i[h^{\ell',i'}]. \tag{16}$$

It trivially follows that the coefficient $\bar{\boldsymbol{\alpha}}$ given by Assumption 1 can be written as $\bar{\boldsymbol{\alpha}} = \boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\mathcal{E}}}$.

We will find the following lemmas useful. Our first two lemmas bound the gap between the empirical and population versions of $\boldsymbol{\Sigma}$ (the left-hand side of the linear system) and $\boldsymbol{\mathcal{E}}$ (the right-hand side of the linear system).

Lemma 3 (Confidence bound for $\boldsymbol{\Sigma}$)**.** *Fix $\delta \in (0,1)$. With probability at least $1-\delta$ over draw of S^{tr} from μ,*

$$|\Sigma_{(\ell,i),(\ell',i')} - \widehat{\Sigma}_{(\ell,i),(\ell',i')}| \leq \mathcal{O}\left(\sqrt{\frac{p_{\ell,i}\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right),$$

where $p_{\ell,i} = \mathbf{E}_{(x,y)\sim\mu}[\phi^{\ell}(x)\mathbf{1}(y=i)]$, and consequently,

$$\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\| \leq \mathcal{O}\left(\sqrt{\frac{L^2 m\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right).$$

Proof. Each row of $\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}$ contains the difference between the elements $\Phi^{\mu,\ell}_i[h]$ and $\widehat{\Phi}^{\text{tr},\ell}_i[h]$ for a classifier h chosen from $\mathcal{H}$. Using multiplicative Chernoff bounds, we have for a fixed h, with probability at least $1-\delta$ over draw of S^{tr} from μ

$$|\Phi^{\mu,\ell}_i[h] - \widehat{\Phi}^{\text{tr},\ell}_i[h]| \leq \mathcal{O}\left(\sqrt{\frac{p_{\ell,i}\log(1/\delta)}{n^{\text{tr}}}}\right),$$

where $p_{\ell,i} = \mathbf{E}_{(x,y)\sim\mu}[\phi^{\ell}(x)\mathbf{1}(y=i)]$. Taking a union bound over all $h \in \mathcal{H}$, we have with probability at least $1-\delta$ over draw of S^{tr} from μ, for any $h \in \mathcal{H}$:

$$|\Phi^{\mu,\ell}_i[h] - \widehat{\Phi}^{\text{tr},\ell}_i[h]| \leq \mathcal{O}\left(\sqrt{\frac{p_{\ell,i}\log(|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right).$$

Taking a union bound over all $Lm \times Lm$ entries, we have with probability at least $1-\delta$, for all $(\ell,i),(\ell',i')$:

$$|\Sigma_{(\ell,i),(\ell',i')} - \widehat{\Sigma}_{(\ell,i),(\ell',i')}| \leq \mathcal{O}\left(\sqrt{\frac{p_{\ell,i}\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right).$$

Upper bounding the operator norm of $\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}$ with the Frobenius norm, we have

$$\begin{aligned}\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\| &\leq \mathcal{O}\left(\sqrt{\frac{\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\sqrt{\sum_{(\ell,i),(\ell',i')} p_{\ell',i'}}\right)\\ &\leq \mathcal{O}\left(\sqrt{\frac{\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\sqrt{\sum_{\ell,i,\ell'}(1)}\right) \leq \mathcal{O}\left(\sqrt{\frac{L^2 m\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right),\end{aligned}$$

where the second inequality uses the fact that $\sum_{i'} p_{\ell',i'} = \mathbf{E}_{x\sim\mathbf{P}^{\mu}}\left[\phi^{\ell'}(x)\right] \leq 1$. □

Lemma 4 (Confidence bound for $\boldsymbol{\mathcal{E}}$)**.** *Fix $\delta \in (0,1)$. With probability at least $1-\delta$ over draw of S^{val} from D,*

$$\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\| \leq \mathcal{O}\left(\sqrt{\frac{Lm\log(Lm/\delta)}{n^{\text{val}}}}\right).$$

Proof. From an application of Hoeffding's inequality, we have for any fixed $h^{\ell,i}$:

$$|\mathcal{E}_{(\ell,i)} - \widehat{\mathcal{E}}_{(\ell,i)}| = |\mathcal{E}^D[h^{\ell,i}] - \widehat{\mathcal{E}}^{\text{val}}[h^{\ell,i}]| = \left|\sum_i \beta_i C^D_{ii}[h^{\ell,i}] - \sum_i \beta_i \widehat{C}^{\text{val}}_{ii}[h^{\ell,i}]\right| \leq \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n^{\text{val}}}}\right),$$

which holds with probability at least $1-\delta$ over draw of S^{val} and uses the fact that each β_i and $C^D_{ii}[h]$ is bounded. Taking a union bound over all Lm probing classifiers, we have:

$$\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\| \leq \mathcal{O}\left(\sqrt{Lm}\sqrt{\frac{\log(Lm/\delta)}{n^{\text{val}}}}\right).$$

Note that we do not need a uniform convergence argument like in Lemma 3 as the probing classifiers are chosen independent of the validation sample. □

Our last two lemmas show that $\mathbf{\Sigma}$ is well-conditioned. We first show that because the probing classifiers $h^{\ell,i}$'s are chosen to satisfy (11), the diagonal and off-diagonal entries of $\mathbf{\Sigma}$ can be lower and upper bounded respectively as follows.

Lemma 5 (Bounds on diagonal and off-diagonal entries of $\mathbf{\Sigma}$)**.** *Fix $\delta \in (0,1)$. With probability at least $1-\delta$ over draw of S^{tr} from μ,*

$$\Sigma_{(\ell,i),(\ell,i)} \geq \gamma - \mathcal{O}\left(\sqrt{\frac{p_{\ell,i}\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right), \forall(\ell,i)$$

and

$$\Sigma_{(\ell,i),(\ell',i')} \leq \omega + \mathcal{O}\left(\sqrt{\frac{p_{\ell,i}\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right), \forall(\ell,i) \neq (\ell',i'),$$

where $p_{\ell,i} = \mathbf{E}_{(x,y)\sim\mu}[\phi^\ell(x)\mathbf{1}(y=i)]$.

Proof. Because the probing classifiers $h^{\ell,i}$'s are chosen from $\mathcal{H}$ to satisfy (11), we have $\widehat{\Sigma}_{(\ell,i),(\ell,i)} \geq \gamma, \forall(\ell,i)$ and $\widehat{\Sigma}_{(\ell,i),(\ell',i')} \leq \omega, \forall(\ell,i) \neq (\ell',i')$. The proof follows from generalization bounds similar to Lemma 3. □

The bounds on the diagonal and off-diagonal entries of $\mathbf{\Sigma}$ then allow us to bound its smallest and largest singular values.

Lemma 6 (Bounds on singular values of $\mathbf{\Sigma}$)**.** *We have $\|\mathbf{\Sigma}\| \leq L\sqrt{m}$. Fix $\delta \in (0,1)$. Suppose $\gamma > 2\sqrt{2}Lm\omega$ and $n^{\text{tr}} \geq \frac{L^2 m\log(Lm|\mathcal{H}|/\delta)}{(\frac{\gamma}{2}-\sqrt{2}Lm\omega)^2}$. With probability at least $1-\delta$ over draw of S^{tr} from μ, $\|\mathbf{\Sigma}^{-1}\| \leq \mathcal{O}\left(\frac{1}{\gamma}\right)$.*

Proof. We first derive a straight-forward upper bound on the the operator norm of $\mathbf{\Sigma}$ in terms of its Frobenius norm:

$$\|\mathbf{\Sigma}\| \leq \sqrt{\sum_{(\ell,i),(\ell',i')}\Sigma^2_{(\ell,i),(\ell',i')}} \leq \sqrt{\sum_{(\ell,i),(\ell',i')}p^2_{\ell',i'}} \leq \sqrt{\sum_{(\ell,i),(\ell',i')}p_{\ell',i'}} \leq \sqrt{\sum_{\ell,i,\ell'}1} = L\sqrt{m},$$

where $p_{\ell,i} = \mathbf{E}_{(x,y)\sim\mu}[\phi^\ell(x)\mathbf{1}(y=i)]$ and the last inequality uses the fact that $\sum_{i'} p_{\ell',i'} = \mathbf{E}_{x\sim\mathbf{P}^\mu}\left[\phi^{\ell'}(x)\right] \leq 1$.

To bound the operator norm of $\|\mathbf{\Sigma}^{-1}\|$, denote $\upsilon_{\ell,i} = \mathcal{O}\left(\sqrt{\frac{p_{\ell,i}\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right)$. From Lemma 5, we can express $\mathbf{\Sigma}$ as a sum of a matrix $\mathbf{A}$ and a diagonal matrix $\mathbf{D}$, i.e. $\mathbf{\Sigma} = \mathbf{A}+\mathbf{D}$, where each $A_{(\ell,i),(\ell,i)} = 0$, $A_{(\ell,i),(\ell',i')} \leq \omega+\upsilon_{\ell,i}, \forall(\ell,i) \neq (\ell',i')$ and $D_{(\ell,i),(\ell,i)} \geq \gamma - \upsilon_{\ell,i}$. Let $\sigma_{\ell,i}(\mathbf{\Sigma})$ denote the (ℓ,i)-th largest singular value of $\mathbf{\Sigma}$. By Weyl's inequality, we have that the singular values of $\mathbf{\Sigma}$ can be bounded in terms of the singular values $\mathbf{D}$ (see e.g., Stewart (1998)):

$$|\sigma_{\ell,i}(\mathbf{\Sigma}) - \sigma_{\ell,i}(\mathbf{D})| \leq \|\mathbf{A}\|,$$

or

$$\sigma_{\ell,i}(\mathbf{D}) - \sigma_{\ell,i}(\mathbf{\Sigma}) \leq \|\mathbf{A}\|.$$

We further have:

$$\begin{aligned}\sigma_{\ell,i}(\mathbf{D}) - \sigma_{\ell,i}(\mathbf{\Sigma}) \;&\leq\; \|\mathbf{A}\| \leq \sqrt{\sum_{(\ell,i)\neq(\ell',i')}(\omega+\upsilon_{\ell,i})^2} + \upsilon_{\ell,i} \leq \sqrt{2}\sqrt{\sum_{(\ell,i)\neq(\ell',i')}\omega^2 + \sum_{(\ell,i)\neq(\ell',i')}\upsilon^2_{\ell,i}} + \upsilon_{\ell,i}\\ &\leq\; \sqrt{2}\sqrt{\sum_{(\ell,i)\neq(\ell',i')}\omega^2} + \sqrt{2}\sqrt{\sum_{(\ell,i)\neq(\ell',i')}\upsilon^2_{\ell,i}}\end{aligned}$$

$$
\begin{aligned}
&\leq \quad \sqrt{2}Lm\omega \,+\, \mathcal{O}\left(\sqrt{\frac{\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right)\sqrt{\sum_{(\ell,i)\neq(\ell',i')} p_{\ell,i}}\\
&\leq \quad \sqrt{2}Lm\omega \,+\, \mathcal{O}\left(\sqrt{\frac{L^2 m\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right).
\end{aligned}
$$

Since $\sigma_{\ell,i}(\mathbf{D}) \geq \gamma - \max_{\ell,i} \upsilon_{\ell,i}$, and

$$
\sigma_{\ell,i}(\mathbf{\Sigma}) \,\geq\, \gamma \,-\, \sqrt{2}Lm\omega \,-\, \mathcal{O}\left(\sqrt{\frac{L^2 m\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right) - \max_{\ell,i} \upsilon_{\ell,i}.
$$

Substituting for $\max_{\ell,i} \upsilon_{\ell,i} \leq \mathcal{O}\left(\sqrt{\frac{\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right)$, and denoting $\xi = \sqrt{2}Lm\omega \,+\, \mathcal{O}\left(\sqrt{\frac{L^2 m\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}}\right)$, we have $\sigma_{\ell,i}(\mathbf{\Sigma}) \,\geq\, \xi$. With this, we can bound operator norm of $\|\mathbf{\Sigma}^{-1}\|$ as:

$$
\|\mathbf{\Sigma}^{-1}\| = \frac{1}{\min_{\ell,i}\sigma_{\ell,i}(\mathbf{\Sigma})} \,\leq\, \frac{1}{\gamma - \xi} \,\leq\, \mathcal{O}\left(\frac{1}{\gamma}\right),
$$

where the last inequality follows from the assumption that $n^{\text{tr}} \geq \frac{L^2 m\log(Lm|\mathcal{H}|/\delta)}{(\frac{\gamma}{2} - \sqrt{2}Lm\omega)^2}$ and hence $\xi \leq \mathcal{O}\left(\gamma/2\right)$. □

We are now ready to prove Theorem 1.

Proof of Theorem 1. The solution from Algorithm 1 is given by $\widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{\Sigma}}^{-1}\widehat{\boldsymbol{\mathcal{E}}}$. Recall we can write the "true" coefficients by $\bar{\boldsymbol{\alpha}} = \mathbf{\Sigma}^{-1}\bar{\boldsymbol{\mathcal{E}}}$, where $\bar{\boldsymbol{\mathcal{E}}}$ is defined in (16), and we also defined $\boldsymbol{\alpha} = \mathbf{\Sigma}^{-1}\boldsymbol{\mathcal{E}}$. The left-hand side of Theorem 1 can then be expanded as:

$$
\begin{aligned}
\|\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\| \quad &\leq \quad \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| + \|\boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\|\\
&\leq \quad \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| + \|\mathbf{\Sigma}^{-1}(\boldsymbol{\mathcal{E}} - \bar{\boldsymbol{\mathcal{E}}})\|\\
&\leq \quad \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| + \|\mathbf{\Sigma}^{-1}\|\|(\boldsymbol{\mathcal{E}} - \bar{\boldsymbol{\mathcal{E}}})\|\\
&\leq \quad \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| + \nu\sqrt{Lm}\|\mathbf{\Sigma}^{-1}\| &(17)\\
&\leq \quad \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| \,+\, \frac{2\nu\sqrt{Lm}}{\gamma}. &(18)
\end{aligned}
$$

Here the second-last step follows from Assumption 1, in particular from $\left|\sum_{\ell,i} \bar{\alpha}^\ell_i \Phi^{\mu,\ell}_i[h] - \mathcal{E}^D[h]\right| \,\leq\, \nu, \forall h$, which gives us that $\left|\sum_{\ell,i} \bar{\alpha}^\ell_i \Phi^{\mu,\ell}_i[h^{\ell',i'}] - \mathcal{E}^D[h^{\ell',i'}]\right| \,\leq\, \nu$, for all ℓ', i'. The last step follows from Lemma 6 and holds with probability at least $1-\delta$ over draw of S^{tr}.

All that remains is to bound the term $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|$. Given that $\widehat{\boldsymbol{\alpha}} = \widehat{\mathbf{\Sigma}}^{-1}\widehat{\boldsymbol{\mathcal{E}}}$. and $\boldsymbol{\alpha} = \mathbf{\Sigma}^{-1}\boldsymbol{\mathcal{E}}$, we can use standard error analysis for linear systems (see e.g., Demmel (1997)) to bound:

$$
\begin{aligned}
\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| \quad &\leq \quad \|\boldsymbol{\alpha}\|\|\mathbf{\Sigma}\|\|\mathbf{\Sigma}^{-1}\|\left(\frac{\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\|}{\|\mathbf{\Sigma}\|} \,+\, \frac{\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\|}{\|\boldsymbol{\mathcal{E}}\|}\right)\\
&\leq \quad \|\mathbf{\Sigma}^{-1}\|^2\|\boldsymbol{\mathcal{E}}\|\left(\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\| \,+\, \|\mathbf{\Sigma}\|\frac{\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\|}{\|\boldsymbol{\mathcal{E}}\|}\right) \;\; \text{(from } \boldsymbol{\alpha} = \mathbf{\Sigma}^{-1}\boldsymbol{\mathcal{E}}\text{)}\\
&\leq \quad \|\mathbf{\Sigma}^{-1}\|^2\|\boldsymbol{\mathcal{E}}\|\left(\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\| \,+\, L\sqrt{m}\frac{\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\|}{\|\boldsymbol{\mathcal{E}}\|}\right) \;\; \text{(from Lemma 6)}\\
&\leq \quad \|\mathbf{\Sigma}^{-1}\|^2\sqrt{Lm}\left(\|\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}}\| \,+\, \frac{L\sqrt{m}}{\sqrt{Lmc}}\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\|\right) \;\; \text{(using } \mathcal{E}_{(\ell,i)} \in (c, 1]\text{)}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \; \|\boldsymbol{\Sigma}^{-1}\|^2\sqrt{Lm}\left(\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\| \;+\; \frac{\sqrt{L}}{c}\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\|\right)\\
&\leq \; \mathcal{O}\left(\frac{\sqrt{Lm}}{\gamma^2}\left(\sqrt{\frac{L^2 m\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}} \;+\; \frac{\sqrt{L}}{c}\sqrt{\frac{Lm\log(Lm/\delta)}{n^{\text{val}}}}\right)\right)\\
&= \; \mathcal{O}\left(\frac{Lm}{\gamma^2}\left(\sqrt{\frac{L\log(Lm|\mathcal{H}|/\delta)}{n^{\text{tr}}}} \;+\; \frac{1}{c}\sqrt{\frac{L\log(Lm/\delta)}{n^{\text{val}}}}\right)\right),
\end{aligned}
$$

where the last two steps follow from Lemmas 3–4 and Lemma 6, and hold with probability at least $1-\delta$ over draws of S^{tr} and S^{val}. Plugging this back into (18) completes the proof. □

B.2. Error Bound for PI-EW

When the metric is linear, we have the following bound on the gap between the metric value achieved by classifier $\widehat{h}$ output by Algorithm 2, and the optimal value. This result will then be useful in proving an error bound for Algorithm 3 in the next section.

Lemma 7 (Error Bound for PI-EW). *Let the input metric be of the form $\widehat{\mathcal{E}}^{\text{lin}}[h] = \sum_i \beta_i \widehat{C}^{\text{val}}_{ii}[h]$ for some (unknown) coefficients $\boldsymbol{\beta} \in \mathbb{R}^m_+$, $\|\boldsymbol{\beta}\| \leq 1$, and denote $\mathcal{E}^{\text{lin}}[h] = \sum_i \beta_i C^D_{ii}[h]$. Let $\bar{\boldsymbol{\alpha}}$ be the associated weighting coefficient for $\mathcal{E}^{\text{lin}}$ in Assumption 1, with $\|\bar{\boldsymbol{\alpha}}\|_1 \leq B$ and with slack ν. Fix $\delta > 0$. Suppose w.p. $\geq 1-\delta$ over draw of S^{tr} and S^{val}, the weight elicitation routine in line 2 of Algorithm 2 provides coefficients $\widehat{\boldsymbol{\alpha}}$ with $\|\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\| \leq \kappa(\delta, n^{\text{tr}}, n^{\text{val}})$, for some function $\kappa(\cdot) > 0$. Let $B' = B + \sqrt{Lm}\,\kappa(\delta, n^{\text{tr}}, n^{\text{val}})$. Then with the same probability, the classifier $\widehat{h}$ output by Algorithm 2 satisfies:*

$$
\max_h \mathcal{E}^{\text{lin}}[h] - \mathcal{E}^{\text{lin}}[\widehat{h}] \quad \leq B'\mathbf{E}_x\left[\|\eta^{\text{tr}}(x) - \widehat{\eta}^{\text{tr}}(x)\|_1\right] \;+\; 2\sqrt{Lm}\,\kappa(\delta, n^{\text{tr}}, n^{\text{val}}) \;+\; 2\nu,
$$

where $\eta^{\text{tr}}_i(x) = \mathbf{P}^\mu(y = i|x)$. Furthermore, when the metric coefficients $\|\boldsymbol{\beta}\| \leq Q$, for some $Q > 0$, then

$$
\max_h \mathcal{E}^{\text{lin}}[h] - \mathcal{E}^{\text{lin}}[\widehat{h}] \quad \leq Q\left(B'\mathbf{E}_x\left[\|\eta^{\text{tr}}(x) - \widehat{\eta}^{\text{tr}}(x)\|_1\right] \;+\; 2\sqrt{Lm}\,\kappa(\delta, n^{\text{tr}}, n^{\text{val}}) \;+\; 2\nu\right).
$$

Proof. For the proof, we will treat $\widehat{h}$ as a classifier that outputs one-hot labels, i.e. as classifier $\widehat{h} : \mathcal{X} \to \{0,1\}^m$ with

$$
\widehat{h}(x) \;=\; \text{onehot}\left(\text{argmax}^*_{i\in[m]}\, \widehat{W}_i(x)\widehat{\eta}^{\text{tr}}_i(x)\right), \tag{19}
$$

where argmax^* breaks ties in favor of the largest class.

Let $\bar{W}_i(x) = \sum_{\ell=1}^L \bar{\alpha}^\ell_i \phi^\ell(x)$ and $\widehat{W}_i(x) = \sum_{\ell=1}^L \widehat{\alpha}^\ell_i \phi^\ell(x)$. It is easy to see that

$$
|\bar{W}_i(x) - \widehat{W}_i(x)| \leq \|\bar{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}\|\sqrt{\sum_{\ell=1}^L \phi^\ell(x)^2} \leq \sqrt{Lm}\|\bar{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\alpha}}\| \leq \sqrt{Lm}\kappa, \tag{20}
$$

where in the second inequality we use $|\phi^\ell(x)| \leq 1$, and in the last inequality, we have shortened the notation $\kappa(\delta, n^{\text{tr}}, n^{\text{val}})$ to κ and for simplicity will avoid mentioning that this holds with high probability.

Further, recall from Assumption 1 that

$$
|\bar{W}_i(x)| \leq \|\bar{\boldsymbol{\alpha}}\|_1 \max_\ell |\phi^\ell(x)| \leq B(1) = B
$$

and so from (20),

$$
|\widehat{W}_i(x)| \leq B + \sqrt{Lm}\kappa. \tag{21}
$$

We also have from Assumption 1 that

$$
\left|\mathcal{E}^{\text{lin}}[h] \;-\; \mathbf{E}_{(x,y)\sim\mu}\left[\sum_{i=1}^m \bar{W}_i(x)\mathbf{1}(y = i)h_i(x)\right]\right| \leq \nu, \forall h.
$$

Equivalently, this can be re-written in terms of the conditional class probabilities $\eta^{\text{tr}}(x) = \mathbf{P}^{\mu}(y = 1|x)$:

$$\left|\mathcal{E}^{\text{lin}}[h] \,-\, \mathbf{E}_{x\sim\mathbf{P}^{\mu}}\left[\sum_{i=1}^{m} \bar{W}_i(x)\eta_i^{\text{tr}}(x)h_i(x)\right]\right| \le \nu, \forall h, \tag{22}$$

where $\mathbf{P}^{\mu}$ denotes the marginal distribution of μ over $\mathcal{X}$. Denoting $h^* \in \operatorname{argmax}_h \, \mathcal{E}^{\text{lin}}[h]$, we then have from (22),

$$\begin{aligned}
&\max_h \mathcal{E}^{\text{lin}}[h] \,-\, \mathcal{E}^{\text{lin}}[\widehat{h}] \\
&= \sum_{i=1}^{m} \mathbf{E}_x\left[\bar{W}_i(x)\eta_i^{\text{tr}}(x)h_i^*(x))\right] \,-\, \sum_{i=1}^{m} \mathbf{E}_x\left[\bar{W}_i(x)\eta_i^{\text{tr}}(x)\widehat{h}_i(x)\right] \,+\, 2\nu \\
&\le \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)\eta_i^{\text{tr}}(x)h_i^*(x))\right] \,-\, \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)\eta_i^{\text{tr}}(x)\widehat{h}_i(x)\right] \,+\, 2\nu \,+\, 2\sqrt{Lm}\kappa \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{(from (20), } \textstyle\sum_{i=1}^{m} \eta_i^{\text{tr}}(x) = 1 \text{ and } h_i(x) \le 1\text{)} \\
&\le \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)\eta_i^{\text{tr}}(x)h_i^*(x))\right] \,-\, \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)\widehat{\eta}_i^{\text{tr}}(x)h_i^*(x))\right] \\
&\qquad\qquad + \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)\widehat{\eta}_i^{\text{tr}}(x)h_i^*(x))\right] \,-\, \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)\eta_i^{\text{tr}}(x)\widehat{h}_i(x)\right] \,+\, 2\nu \,+\, 2\sqrt{Lm}\kappa
\end{aligned}$$

From the definition of $\widehat{h}$ in (19), we have that $\sum_{i=1}^{m} \widehat{W}_i(x)\widehat{\eta}_i^{\text{tr}}(x)\widehat{h}_i(x) \ge \sum_{i=1}^{m} \widehat{W}_i(x)\widehat{\eta}_i^{\text{tr}}(x)h_i(x)$, for all $h : \mathcal{X} \to \Delta_m$. Therefore,

$$\begin{aligned}
&\max_h \mathcal{E}^{\text{lin}}[h] \,-\, \mathcal{E}^{\text{lin}}[\widehat{h}] \\
&\le \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)\eta_i^{\text{tr}}(x)h_i^*(x))\right] \,-\, \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)\widehat{\eta}_i^{\text{tr}}(x)h_i^*(x))\right] \\
&\qquad\qquad + \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)\widehat{\eta}_i^{\text{tr}}(x)\widehat{h}_i(x))\right] \,-\, \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)\eta_i^{\text{tr}}(x)\widehat{h}_i(x)\right] \,+\, 2\nu \,+\, 2\sqrt{Lm}\kappa \\
&\le \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)|\eta_i^{\text{tr}}(x) - \widehat{\eta}_i^{\text{tr}}(x)|h_i^*(x)\right] \,+\, \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)|\eta_i^{\text{tr}}(x) - \widehat{\eta}_i^{\text{tr}}(x)|\widehat{h}_i(x)\right] \,+\, 2\nu \,+\, 2\sqrt{Lm}\kappa \\
&\le \sum_{i=1}^{m} \mathbf{E}_x\left[\widehat{W}_i(x)|\eta_i^{\text{tr}}(x) - \widehat{\eta}_i^{\text{tr}}(x)||h_i^*(x) - \widehat{h}_i(x)|\right] \,+\, 2\nu \,+\, 2\sqrt{Lm}\kappa \\
&\le \mathbf{E}_x\left[\max_i \left(\widehat{W}_i(x)|h_i^*(x) - \widehat{h}_i(x)|\right) \|\eta(x) - \widehat{\eta}(x)\|_1\right] \,+\, 2\nu \,+\, 2\sqrt{Lm}\kappa \\
&\le (B + \sqrt{Lm}\kappa)\, \mathbf{E}_x\left[\|\eta(x) - \widehat{\eta}(x)\|_1\right] \,+\, 2\nu \,+\, 2\sqrt{Lm}\kappa,
\end{aligned}$$

where the last step follows from (21) and $|h_i(x) - \widehat{h}_i(x)| \le 1$. This completes the proof. The second part, where $\|\boldsymbol{\beta}\| \le Q$, follows by applying Assumption 1 to normalized coefficients $\boldsymbol{\beta}/\|\boldsymbol{\beta}\|$, and scaling the associated slack ν by Q. □

B.3. Proof of Theorem 2

We will make a couple of minor changes to the algorithm to simplify the analysis. Firstly, instead of using the same sample S^{val} for both estimating the example weights (through call to **PI-EW** in line 9) and estimating confusion matrices $\widehat{\mathbf{C}}^{\text{val}}$ (in line 10), we split S^{val} into two halves, use one half for the first step and the other half for the second step. Using independent samples for the two steps, we will be able to derive straight-forward confidence bounds on the estimated confusion matrices in each case. In our experiments however, we find the algorithm to be effective even when a common sample is used for both steps. Secondly, we modify line 8 to include a shifted version of the metric $\widehat{\mathcal{E}}^{\text{val}}$, so that later in Appendix D when we handle the case of "unknown ψ", we can avoid having to keep track of an additive constant in the gradient coefficients.

Algorithm 3***:** **F**rank-**W**olfe with **E**licited **G**radients (**FW-EG**) for General Diagonal Metrics

```
1: Input: $\widehat{\mathcal{E}}^{\text{val}}$, Basis functions $\phi^1, \ldots, \phi^L : \mathcal{X} \to [0,1]$, Pre-trained $\widehat{\eta}^{\text{tr}} : \mathcal{X} \to \Delta_m$, $S^{\text{tr}} \sim \mu$, $S^{\text{val}} \sim D$ split into two halves
   $S_1^{\text{val}}$ and $S_2^{\text{val}}$ of sizes $\lceil n^{\text{val}}/2 \rceil$ and $\lfloor n^{\text{val}}/2 \rfloor$ respectively, $T$, $\epsilon$
2: Initialize classifier $h^0$ and $\mathbf{c}^0 = \text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h^0])$
3: For $t = 0$ to $T-1$ do
4:   if $\mathcal{E}^D[h] = \psi(C_{11}^D[h], \ldots, C_{mm}^D[h])$ for known $\psi$:
5:     $\boldsymbol{\beta}^t \;=\; \nabla\psi(\mathbf{c}^t)$
6:     $\widehat{\mathcal{E}}^{\text{lin}}[h] \;=\; \sum_i \beta_i^t \widehat{C}_{ii}^{\text{val}}[h]$, evaluated using $S_1^{\text{val}}$
7:   else
8:     $\widehat{\mathcal{E}}^{\text{lin}}[h] = \widehat{\mathcal{E}}^{\text{val}}[h] - \widehat{\mathcal{E}}^{\text{val}}[h^t]$, evaluated using $S_1^{\text{val}}$  {small $\epsilon$ recommended}
9:   $\widehat{f} = $ PI-EW$(\widehat{\mathcal{E}}^{\text{lin}}, \phi^1, ..., \phi^L, \widehat{\eta}^{\text{tr}}, S^{\text{tr}}, S_1^{\text{val}}, h^t, \epsilon)$
10:  $\widetilde{\mathbf{c}} = \text{diag}(\widehat{\mathbf{C}}^{\text{val}}[\widehat{f}])$, evaluated using $S_2^{\text{val}}$
11:  $h^{t+1} = \big(1 - \frac{2}{t+1}\big)h^t + \frac{2}{t+1}\text{onehot}(\widehat{f})$
12:  $\mathbf{c}^{t+1} = \big(1 - \frac{2}{t+1}\big)\mathbf{c}^t + \frac{2}{t+1}\widetilde{\mathbf{c}}$
13: End For
14: Output: $\widehat{h} = h^T$
```

Theorem ((Restated) **Error Bound for FW-EG with known** ψ). *Let $\mathcal{E}^D[h] = \psi(C_{11}^D[h], \ldots, C_{mm}^D[h])$ for a* known *concave function $\psi : [0,1]^m \to \mathbb{R}_+$, which is Q-Lipschitz, and λ-smooth w.r.t. the ℓ_1-norm. Let $\widehat{\mathcal{E}}^{\text{val}}[h] = \psi(\widehat{C}_{11}^{\text{val}}[h], \ldots, \widehat{C}_{mm}^{\text{val}}[h])$. Fix $\delta \in (0,1)$. Suppose Assumption 1 holds with slack ν, and for any linear metric $\sum_i \beta_i C_{ii}^D[h]$ with $\|\boldsymbol{\beta}\| \leq 1$, whose associated weight coefficients is $\bar{\boldsymbol{\alpha}}$ with $\|\bar{\boldsymbol{\alpha}}\| \leq B$, w.p. $\geq 1-\delta$ over draw of S^{tr} and S_1^{val}, the weight elicitation routine in Algorithm 1 outputs coefficients $\widehat{\boldsymbol{\alpha}}$ with $\|\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\| \leq \kappa(\delta, n^{\text{tr}}, n^{\text{val}})$, for some function $\kappa(\cdot) > 0$. Let $B' = B + \sqrt{Lm}\,\kappa(\delta/T, n^{\text{tr}}, n^{\text{val}})$. Assume $m \leq n^{\text{val}}$. Then w.p. $\geq 1-\delta$ over draws of S^{tr} and S^{val} from D and μ resp., the classifier $\widehat{h}$ output by Algorithm 3* after T iterations satisfies:*

$$\max_h \mathcal{E}^D[h] - \mathcal{E}^D[\widehat{h}] \;\leq\; 2QB'\mathbf{E}_x\left[\|\eta^{\text{tr}}(x) - \widehat{\eta}^{\text{tr}}(x)\|_1\right] + 4Q\nu + 4Q\sqrt{Lm}\,\kappa(\delta/T, n^{\text{tr}}, n^{\text{val}})$$
$$+\mathcal{O}\left(\lambda m\sqrt{\frac{m\log(n^{\text{val}})\log(m) + \log(m/\delta)}{n^{\text{val}}}} + \frac{\lambda}{T}\right).$$

The proof adapts techniques from Narasimhan et al. (2015b), who show guarantees for a Frank-Wolfe based learning algorithm with a known ψ in the *absence* of distribution shift. The main proof steps are listed below:

- Prove a generalization bound for the confusion matrices $\widehat{\mathbf{C}}^{\text{val}}$ evaluated in line 10 on the validation sample (Lemma 8)
- Establish an error bound for the call to **PI-EW** in line 9 (Lemma 7 in previous section)
- Combine the above two results to show that the classifier $\widehat{f}$ returned in line 9 is an approximate linear maximizer needed by the Frank-Wolfe algorithm (Lemma 9)
- Combine Lemma 9 with a convergence guarantee for the outer Frank-Wolfe algorithm (Narasimhan et al., 2015b; Jaggi, 2013) (using convexity of the space of confusion matrices $\mathcal{C}$) to complete the proof (Lemmas 10–11).

Lemma 8 (Generalization bound for $\mathbf{C}^D$)**.** *Fix $\delta \in (0,1)$. Let $\widehat{\eta}^{\text{tr}} : \mathcal{X} \to \Delta_m$ be a fixed class probability estimator. Let $\mathcal{G} = \{h : \mathcal{X} \to [m] \,|\, h(x) \in \operatorname{argmax}_{i \in [m]} \beta_i \widehat{\eta}_i^{\text{tr}}(x)$ for some $\boldsymbol{\beta} \in \mathbb{R}_+^m\}$ be the set of plug-in classifiers defined with $\widehat{\eta}^{\text{tr}}$. Let*

$$\bar{\mathcal{G}} = \{h(x) = \textstyle\sum_{t=1}^T u_t h_t(x) \,|\, T \in \mathbb{N}, h_1, \ldots, h_T \in \mathcal{G}, \mathbf{u} \in \Delta_T\}$$

be the set of all randomized classifiers constructed from a finite number of plug-in classifiers in $\mathcal{G}$. Assume $m \leq n^{\text{val}}$. Then with probability at least $1-\delta$ over draw of S^{val} from D, then for $h \in \bar{\mathcal{G}}$:

$$\|\mathbf{C}^D[h] - \widehat{\mathbf{C}}^{\text{val}}[h]\|_\infty \leq \mathcal{O}\left(\sqrt{\frac{m\log(m)\log(n^{\text{val}}) + \log(m/\delta)}{n^{\text{val}}}}\right).$$

Proof. The proof follows from standard convergence based generalization arguments, where we bound the capacity of the class of plug-in classifiers $\mathcal{G}$ in terms of its Natarajan dimension (Natarajan, 1989; Daniely et al., 2011). Applying Theorem 21 from (Daniely et al., 2011), we have that the Natarajan dimension of $\mathcal{G}$ is at most $d = m\log(m)$. Applying the generalization bound in Theorem 13 in Daniely et al. (2015), along with the assumption that $m \leq n^{\text{val}}$, we have for any $i \in [m]$, with probability at least $1-\delta$ over draw of S^{val} from D, for any $h \in \mathcal{G}$:

$$|C^D_{ii}[h] - \widehat{C}^{\text{val}}_{ii}[h]| \leq \mathcal{O}\left(\sqrt{\frac{m\log(m)\log(n^{\text{val}}) + \log(1/\delta)}{n^{\text{val}}}}\right).$$

Further note that for any randomized classifier $\bar{h}(x) = \sum_{t=1}^{T} u_t h_t(x) \in \bar{\mathcal{G}}$, for some $\mathbf{u} \in \Delta_T$,

$$|C^D_{ii}[\bar{h}] - \widehat{C}^{\text{val}}_{ii}[\bar{h}]| \leq \sum_{t=1}^{T} u_t |C^D_{ii}[h_t] - \widehat{C}^{\text{val}}_{ii}[h_t]| \leq \mathcal{O}\left(\sqrt{\frac{m\log(m)\log(n^{\text{val}}) + \log(1/\delta)}{n^{\text{val}}}}\right),$$

where the first inequality follows from linearity of expectations. Taking a union bound over all diagonal entries $i \in [m]$ completes the proof. □

We next show that the call to **PI-EW** in line 9 of Algorithm 3 computes an approximate maximizer $\widehat{f}$ for $\widehat{\mathcal{E}}^{\text{lin}}$. This is an extension of Lemma 26 in Narasimhan et al. (2015b).

Lemma 9 (Approximation error in linear maximizer $\widehat{f}$)**.** *For each iteration t in Algorithm 3, denote $\bar{\mathbf{c}}^t = \text{diag}(\mathbf{C}^D[h^t])$, and $\bar{\boldsymbol{\beta}}^t = \nabla\psi(\bar{\mathbf{c}}^t)$. Suppose the assumptions in Theorem 2 hold. Let $B' = B + \sqrt{Lm}\,\kappa(\delta, n^{\text{tr}}, n^{\text{val}})$. Assume $m \leq n^{\text{val}}$. Then w.p. $\geq 1-\delta$ over draw of S^{tr} and S^{val} from μ and D resp., for any $t = 1, \ldots, T$, the classifier $\widehat{f}$ returned by* **PI-EW** *in line 9 satisfies:*

$$\max_h \sum_i \bar{\beta}^t_i C^D_{ii}[h] \,-\, \sum_i \bar{\beta}^t_i C^D_{ii}[\widehat{f}] \,\leq\, QB'\mathbf{E}_x\left[\|\eta^{\text{tr}}(x) - \widehat{\eta}^{\text{tr}}(x)\|_1\right] \,+\, 2Q\nu$$
$$+\, 2Q\sqrt{Lm}\,\kappa\left(\tfrac{\delta}{T}, n^{\text{tr}}, n^{\text{val}}\right) \,+\, \mathcal{O}\left(\lambda m\sqrt{\frac{m\log(m)\log(n^{\text{val}}) + \log(m/\delta)}{n^{\text{val}}}}\right).$$

Proof. The proof uses Theorem 1 to bound the approximation errors in the linear maximizer $\widehat{f}$ (coupled with a union bound over T iterations), and Lemma 8 to bound the estimation errors in the confusion matrix $\mathbf{c}^t$ used to compute the gradient $\nabla\psi(\mathbf{c}^t)$.

Recall from Algorithm 3 that $\mathbf{c}^t = \text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h^t])$ and $\boldsymbol{\beta}^t = \nabla\psi(\mathbf{c}^t)$. Note that these are approximations to the actual quantities we are interested in $\bar{\mathbf{c}}^t = \text{diag}(\mathbf{C}^D[h^t])$ and $\bar{\boldsymbol{\beta}}^t = \nabla\psi(\bar{\mathbf{c}}^t)$, both of which are evaluated using the population confusion matrix. Also, $\|\boldsymbol{\beta}\| = \|\nabla\psi(\mathbf{c}^t)\| \leq Q$ from Q-Lipschitzness of ψ.

Fix iteration t, and let $h^* \in \text{argmax}_h \sum_i \bar{\beta}^t_i C^D_{ii}[h]$ for this particular iteration. Then:

$$\begin{aligned}
&\sum_i \bar{\beta}^t_i C^D_{ii}[h^*] - \sum_i \bar{\beta}^t_i C^D_{ii}[\widehat{f}] \\
&= \sum_i \bar{\beta}^t_i C^D_{ii}[h^*] - \sum_i \beta^t_i C^D_{ii}[h^*] + \sum_i \beta^t_i C^D_{ii}[h^*] - \sum_i \beta^t_i C^D_{ii}[\widehat{f}] + \sum_i \beta^t_i C^D_{ii}[\widehat{f}] - \sum_i \bar{\beta}^t_i C^D_{ii}[\widehat{f}] \\
&\leq \|\boldsymbol{\beta}^t - \bar{\boldsymbol{\beta}}^t\|_\infty \sum_i C^D_{ii}[h^*] + \sum_i \beta^t_i C^D_{ii}[h^*] - \sum_i \beta^t_i C^D_{ii}[\widehat{f}] + \|\boldsymbol{\beta}^t - \bar{\boldsymbol{\beta}}^t\|_\infty \sum_i C^D_{ii}[\widehat{f}] \\
&\leq \|\boldsymbol{\beta}^t - \bar{\boldsymbol{\beta}}^t\|_\infty(1) + \max_h \sum_i \beta^t_i C^D_{ii}[h] - \sum_i \beta^t_i C^D_{ii}[\widehat{f}] + \|\boldsymbol{\beta}^t - \bar{\boldsymbol{\beta}}^t\|_\infty(1) \quad (\text{because } \textstyle\sum_{i,j} C^D_{ij}[h] = 1) \\
&= 2\|\boldsymbol{\beta}^t - \bar{\boldsymbol{\beta}}^t\|_\infty + \max_h \sum_i \beta^t_i C^D_{ii}[h] - \sum_i \beta^t_i C^D_{ii}[\widehat{f}] \\
&= 2\|\nabla\psi(\mathbf{c}^t) - \nabla\psi(\bar{\mathbf{c}}^t)\|_\infty + \max_h \sum_i \beta^t_i C^D_{ii}[h] - \sum_i \beta^t_i C^D_{ii}[\widehat{f}]
\end{aligned}$$

$$\leq 2\lambda\|\mathbf{c}^t - \bar{\mathbf{c}}^t\|_1 \,+\, \max_h \sum_i \beta_i^t C_{ii}^D[h] \,-\, \sum_i \beta_i^t C_{ii}^D[\widehat{f}] \quad \text{(because } \psi \text{ is } \lambda\text{-smooth w.r.t. the } \ell_1 \text{ norm)}$$

$$\leq 2\lambda m\|\mathbf{c}^t - \bar{\mathbf{c}}^t\|_\infty \,+\, \max_h \sum_i \beta_i^t C_{ii}^D[h] \,-\, \sum_i \beta_i^t C_{ii}^D[\widehat{f}]$$

$$\leq \mathcal{O}\left(\lambda m\sqrt{\frac{m\log(m)\log(n^{\text{val}}) + \log(m/\delta)}{n^{\text{val}}}}\right) + QB'\mathbf{E}_x\left[\|\eta^{\text{tr}}(x) - \widehat{\eta}^{\text{tr}}(x)\|_1\right] + 2Q\sqrt{Lm}\,\kappa(\delta, n^{\text{tr}}, n^{\text{val}}) + 2Q\nu, \quad (23)$$

where $B' = B + \sqrt{Lm}\,\kappa(\delta, n^{\text{tr}}, n^{\text{val}})$. The last step holds with probability at least $1-\delta$ over draw of S^{val} and S^{tr}, and follows from Lemma 8 and Lemma 7 (using $\|\boldsymbol{\beta}^t\| \leq Q$). The first bound on $\|\mathbf{c}^t - \bar{\mathbf{c}}^t\|_\infty = \|\widehat{\mathbf{C}}^{\text{val}}[h^t] - \mathbf{C}^D[h^t]\|_\infty$ holds for any randomized classifier h^t constructed from a finite number of plug-in classifiers. The second bound on the linear maximization errors holds only for a fixed t, and so we need to take a union bound over all iterations $t = 1, \ldots, T$, to complete the proof. Note that because we use two independent samples S_1^{val} and S_2^{val} for the two bounds, they each hold with high probability over draws of S_1^{val} and S_2^{val} respectively, and hence with high probability over draw of S^{val}. □

Our last two lemmas restate results from Narasimhan et al. (2015b). The first shows convexity of the space of confusion matrices (Proposition 10 from their paper), and the second applies a result from Jaggi (2013) to show convergence of the classical Frank-Wolfe algorithm with approximate linear maximization steps (Theorem 16 in Narasimhan et al. (2015b)).

Lemma 10 (Convexity of space of confusion matrices). *Let $\mathcal{C} = \{\text{diag}(\mathbf{C}^D[h])\,|\,h : \mathcal{X}\to\Delta_m\}$ denote the set of all confusion matrices achieved by some randomized classifier $h : \mathcal{X}\to\Delta_m$. Then $\mathcal{C}$ is convex.*

Proof. For any $\mathbf{C}^1, \mathbf{C}^2 \in \mathcal{C}$, $\exists h_1, h_2 : \mathcal{X}\to\Delta_m$ such that $\mathbf{c}^1 = \text{diag}(\mathbf{C}^D[h_1])$ and $\mathbf{c}^2 = \text{diag}(\mathbf{C}^D[h_2])$. We need to show that for any $u \in [0,1]$, $u\mathbf{c}^1 + (1-u)\mathbf{c}^2 \in \mathcal{C}$. This is true because the randomized classifier $h(x) = uh_1(x) + (1-u)h_2(x)$ yields a confusion matrix $\text{diag}(\mathbf{C}^D[h]) = u\,\text{diag}(\mathbf{C}^D[h_1]) + (1-u)\text{diag}(\mathbf{C}^D[h_2]) = u\mathbf{c}^1 + (1-u)\mathbf{c}^2 \in \mathcal{C}$. □

Lemma 11 (Frank-Wolfe with approximate linear maximization (Narasimhan et al., 2015b)). *Let $\mathcal{E}^D[h] = \psi(C_{11}^D[h], \ldots, C_{mm}^D[h])$ for a concave function $\psi : [0,1]^m\to\mathbb{R}_+$ that is λ-smooth w.r.t. the ℓ_1-norm. For each iteration t, define $\bar{\boldsymbol{\beta}}^t = \nabla\psi(\text{diag}(\mathbf{C}^D[h^t]))$. Suppose line 9 of Algorithm 3 returns a classifier $\widehat{f}$ such that $\max_h \sum_i \bar{\beta}_i^t C_{ii}^D[h] - \sum_i \bar{\beta}_i^t C_{ii}^D[\widehat{f}] \leq \Delta, \forall t \in [T]$. Then the classifier $\widehat{h}$ output by Algorithm 3 after T iterations satisfies:*

$$\max_h \mathcal{E}^D[h] - \mathcal{E}^D[\widehat{h}] \,\leq\, 2\Delta + \frac{8\lambda}{T+2}.$$

Proof of Theorem 2. The proof follows by plugging in the result from Lemma 9 into Lemma 11. □

C. Error Bound for Weight Elicitation with Fixed Probing Classifiers

We first state a general error bound for Algorithm 1 in terms of the singular values of $\mathbf{\Sigma}$ for any *fixed* choices for the probing classifiers. We then bound the singular values for the fixed choices in (12) under some specific assumptions.

Theorem 12 (Error bound on elicited weights with fixed probing classifiers). *Let $\mathcal{E}^D[h] = \sum_i \beta_i C_{ii}^D[h]$ for some (unknown) $\boldsymbol{\beta} \in \mathbb{R}^m$, and let $\widehat{\mathcal{E}}^{\text{val}}[h] = \sum_i \beta_i \widehat{C}_{ii}^{\text{val}}[h]$. Let $\bar{\boldsymbol{\alpha}}$ be the associated coefficient in Assumption 1 for metric $\mathcal{E}^D$. Fix $\delta \in (0,1)$. Then for any fixed choices of the probing classifiers $h^{\ell,i}$, we have with probability $\geq 1-\delta$ over draws of S^{tr} and S^{val} from μ and D resp., the coefficients $\widehat{\boldsymbol{\alpha}}$ output by Algorithm 1 satisfies:*

$$\|\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\| \leq \mathcal{O}\left(\frac{1}{\sigma_{\min}(\mathbf{\Sigma})^2}\left(Lm\sqrt{\frac{L\log(Lm/\delta)}{n^{\text{tr}}}} + \sigma_{\max}(\mathbf{\Sigma})\sqrt{\frac{Lm\log(Lm/\delta)}{n^{\text{val}}}}\right) + \frac{\nu\sqrt{Lm}}{\sigma_{\min}(\mathbf{\Sigma})}\right),$$

where $\sigma_{\min}(\mathbf{\Sigma})$ and $\sigma_{\min}(\mathbf{\Sigma})$ are respectively the smallest and largest singular values of $\mathbf{\Sigma}$.

Proof. The proof follows the same steps as Theorem 1, except for the bound on $\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|$. Specifically, we have from (17):

$$\|\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\| \,\leq\, \|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| \,+\, \nu\sqrt{Lm}\|\mathbf{\Sigma}^{-1}\|. \quad (24)$$

We next bound:

$$\begin{aligned}
\|\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\| &\leq \|\boldsymbol{\alpha}\|\|\boldsymbol{\Sigma}\|\|\boldsymbol{\Sigma}^{-1}\|\left(\frac{\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|}{\|\boldsymbol{\Sigma}\|} + \frac{\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\|}{\|\boldsymbol{\mathcal{E}}\|}\right)\\
&\leq \|\boldsymbol{\Sigma}^{-1}\|^2\|\boldsymbol{\mathcal{E}}\|\left(\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\| + \|\boldsymbol{\Sigma}\|\frac{\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\|}{\|\boldsymbol{\mathcal{E}}\|}\right) \;\; (\text{from } \boldsymbol{\alpha} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mathcal{E}})\\
&\leq \|\boldsymbol{\Sigma}^{-1}\|^2\left(\|\boldsymbol{\mathcal{E}}\|\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\| + \|\boldsymbol{\Sigma}\|\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\|\right)\\
&\leq \|\boldsymbol{\Sigma}^{-1}\|^2\left(\sqrt{Lm}\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\| + \|\boldsymbol{\Sigma}\|\|\boldsymbol{\mathcal{E}} - \widehat{\boldsymbol{\mathcal{E}}}\|\right) \;\; (\text{as } \mathcal{E}^D[h] \in [0,1])\\
&\leq \mathcal{O}\left(\frac{1}{\sigma_{\min}(\boldsymbol{\Sigma})^2}\left(\sqrt{Lm}\sqrt{\frac{L^2 m\log(Lm/\delta)}{n^{\text{tr}}}} + \sigma_{\max}(\boldsymbol{\Sigma})\sqrt{\frac{Lm\log(Lm/\delta)}{n^{\text{val}}}}\right)\right),
\end{aligned}$$

where the last step follows from an adaptation of Lemma 3 (where $\mathcal{H}$ contains the Lm fixed classifiers in (12)) and from Lemma 4. The last statement holds with probability at least $1-\delta$ over draws of S^{tr} and S^{val}. Substituting this bound back in (24) completes the proof. □

We next provide a bound on the singular values of $\boldsymbol{\Sigma}$ for a specialized setting where the the probing classifiers $h^{\ell,i}$ are set to (12), the basis functions ϕ^{ℓ}'s divide the data into disjoint clusters, and the base classifier $\bar{h}$ is close to having "uniform accuracies" across all the clusters and classes.

Lemma 13. *Let $h^{\ell,i}$'s be defined as in (12). Suppose for any x, $\phi^{\ell}(x) \in \{0,1\}$ and $\phi^{\ell}(x)\phi^{\ell'}(x) = 0, \forall \ell \neq \ell'$. Let $p_{\ell,i} = \mathbf{E}_{(x,y)\sim\mu}[\phi^{\ell}(x)\mathbf{1}(y=i)]$. Let $\bar{h}$ be such that $\kappa - \tau \leq \Phi_i^{\mu,\ell}[\bar{h}] \leq \kappa, \forall \ell, i$ and for some $\kappa < \frac{1}{m}$ and $\tau < \kappa$. Then:*

$$\sigma_{\max}(\boldsymbol{\Sigma}) \leq L\max_{\ell,i} p_{\ell,i} + \Delta; \quad \sigma_{\min}(\boldsymbol{\Sigma}) \geq \epsilon(1-m\kappa)\min_{\ell,i} p_{\ell,i} - \Delta,$$

where $\Delta = Lm\tau\max_{\ell,i} p_{\ell,i}$.

Proof. We first write the matrix $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \bar{\boldsymbol{\Sigma}} + \mathbf{E}$, where

$$\bar{\boldsymbol{\Sigma}} = \begin{bmatrix}
p_{1,1}(\epsilon + (1-\epsilon)\kappa) & p_{1,2}(1-\epsilon)\kappa & \dots & p_{1,m}(1-\epsilon)\kappa & p_{2,1}\kappa & \dots & p_{L,m}\kappa\\
p_{1,1}(1-\epsilon)\kappa & p_{1,2}(\epsilon + (1-\epsilon)\kappa) & \dots & p_{1,m}(1-\epsilon)\kappa & p_{2,1}\kappa & \dots & p_{L,m}\kappa\\
 & & & \vdots & & & \\
p_{1,1}\kappa & p_{1,2}\kappa & \dots & p_{1,m}\kappa & p_{2,1}\kappa & \dots & p_{L,m}(\epsilon + (1-\epsilon)\kappa)
\end{bmatrix},$$

and $\mathbf{E} \in \mathbb{R}^{Lm\times Lm}$ with each $|E_{(\ell,i),(\ell',i')}| \leq \max_{\ell,i} p_{\ell,i}\left(\kappa - \Phi_i^{\mu,\ell}[\bar{h}]\right) \leq \tau\max_{\ell,i} p_{\ell,i}$.

The matrix $\bar{\boldsymbol{\Sigma}}$ can in turn be written as a product of a *symmetric* matrix $\mathbf{A} \in \mathbb{R}^{Lm\times Lm}$ and a *diagonal* matrix $\mathbf{D} \in \mathbb{R}^{Lm\times Lm}$:

$$\bar{\boldsymbol{\Sigma}} = \mathbf{A}\mathbf{D},$$

where

$$\mathbf{A} = \begin{bmatrix}
\epsilon + (1-\epsilon)\kappa & (1-\epsilon)\kappa & \dots & (1-\epsilon)\kappa & \kappa & \dots & \kappa\\
(1-\epsilon)\kappa & \epsilon + (1-\epsilon)\kappa & \dots & (1-\epsilon)\kappa & \kappa & \dots & \kappa\\
 & & & \vdots & & & \\
(1-\epsilon)\kappa & (1-\epsilon)\kappa & \dots & \epsilon + (1-\epsilon)\kappa & \kappa & \dots & \kappa\\
 & & & \vdots & & & \\
\kappa & \kappa & \dots & \kappa & \epsilon + (1-\epsilon)\kappa & \dots & (1-\epsilon)\kappa\\
 & & & \vdots & & & \\
\kappa & \kappa & \dots & \kappa & (1-\epsilon)\kappa & \dots & \epsilon + (1-\epsilon)\kappa
\end{bmatrix}; \quad \mathbf{D} = \text{diag}(p_{1,1}, \dots, p_{L,m}).$$

We can then bound the largest and smallest singular values of $\mathbf{\Sigma}$ in terms of those of $\mathbf{A}$ and $\mathbf{D}$. Using Weyl's inequality (see e.g., (Stewart, 1998)), we have

$$\sigma_{\max}(\mathbf{\Sigma}) \leq \sigma_{\max}(\bar{\mathbf{\Sigma}}) + \|\mathbf{E}\| \leq \|\mathbf{A}\|\|\mathbf{D}\| \,+\, \|\mathbf{E}\| = \sigma_{\max}(\mathbf{A})\sigma_{\max}(\mathbf{D}) + \|\mathbf{E}\|.$$

and

$$\sigma_{\min}(\mathbf{\Sigma}) \geq \sigma_{\min}(\bar{\mathbf{\Sigma}}) \,-\, \|\mathbf{E}\| = \frac{1}{\|\bar{\mathbf{\Sigma}}^{-1}\|} \,-\, \|\mathbf{E}\| \geq \frac{1}{\|\mathbf{A}^{-1}\|\|\mathbf{D}^{-1}\|} \,-\, \|\mathbf{E}\| = \sigma_{\min}(\mathbf{A})\sigma_{\min}(\mathbf{D}) \,-\, \|\mathbf{E}\|.$$

Further, we have $\|\mathbf{E}\| \leq \|\mathbf{E}\|_F \leq Lm\tau \max_{\ell,i} p_{\ell,i} = \Delta$, giving us:

$$\sigma_{\max}(\mathbf{\Sigma}) \leq \sigma_{\max}(\mathbf{A})\sigma_{\max}(\mathbf{D}) + \Delta. \tag{25}$$

$$\sigma_{\min}(\mathbf{\Sigma}) \geq \sigma_{\min}(\mathbf{A})\sigma_{\min}(\mathbf{D}) - \Delta. \tag{26}$$

All that remains is to bound the singular values of $\mathbf{\Sigma}$ and $\mathbf{D}$. Since $\mathbf{D}$ is a diagonal matrix, it's singular values are given by its diagonal entries:

$$\sigma_{\max}(\mathbf{D}) = \max_{\ell,i}\, p_{\ell,i}; \quad \sigma_{\min}(\mathbf{D}) = \min_{\ell,i}\, p_{\ell,i}.$$

The matrix $\mathbf{A}$ is symmetric and has a certain block structure. It's singular values are the same as the positive magnitudes of its Eigen values. We first write out it's Lm Eigen vectors:

$$\begin{array}{rlccccl}
\mathbf{x}^{1,1} & = [& \overbrace{1,-1,0,\ldots,0,}^{m \text{ entries}} & \overbrace{0,\ldots,0,}^{m \text{ entries}} & \ldots & \overbrace{0,\ldots,0}^{m \text{ entries}} &] \\
\mathbf{x}^{1,2} & = [& 1,0,-1,\ldots,0, & 0,\ldots,0, & \ldots & 0,\ldots,0 &] \\
 & & & \vdots & & & \\
\mathbf{x}^{1,m-1} & = [& 1,0,0,\ldots,-1, & 0,\ldots,0, & \ldots & 0,\ldots,0 &] \\
\mathbf{x}^{1,m} & = [& 1,\ldots,1, & -1,\ldots,-1, & \ldots & 0,\ldots,0 &] \\
\mathbf{x}^{2,1} & = [& 0,\ldots,0, & 1,-1,0,\ldots,0, & \ldots & 0,\ldots,0 &] \\
 & & & \vdots & & & \\
\mathbf{x}^{2,m-1} & = [& 0,\ldots,0, & 1,0,0,\ldots,-1, & \ldots & 0,\ldots,0 &] \\
\mathbf{x}^{2,m} & = [& -1,\ldots,-1, & 1,\ldots,1, & \ldots & 0,\ldots,0 &] \\
 & & & \vdots & & & \\
\mathbf{x}^{L,1} & = [& 0,\ldots,0, & 0,\ldots,0, & \ldots & 1,-1,0,\ldots,0 &] \\
 & & & \vdots & & & \\
\mathbf{x}^{L,m-1} & = [& 0,\ldots,0, & 0,\ldots,0, & \ldots & 1,0,0,\ldots,-1 &] \\
\mathbf{x}^{L,m} & = [& 1,\ldots,1, & 1,\ldots,1, & \ldots & 1,\ldots,1 &]
\end{array}$$

One can then verify that the Lm Eigen values of $\mathbf{A}$ are ϵ with a multiplicity of $(L-1)m$, $\epsilon(1-m\kappa)$ with a multiplicity of $m-1$ and $(L-\epsilon)m\kappa+\epsilon$ with a multiplicity of 1. Therefore:

$$\sigma_{\max}(\mathbf{A}) \leq L; \quad \sigma_{\min}(\mathbf{A}) = \epsilon(1-m\kappa).$$

Substituting the singular (Eigen) values of $\mathbf{A}$ and $\mathbf{D}$ into (25) and (26) completes the proof. □

In the above lemma, the base classifier $\bar{h}$ is assumed to have roughly uniformly low accuracies for all classes and clusters, and the closer it is to having uniform accuracies, i.e. the smaller the value of τ, the tighter are the bounds.

We have shown a bound on the singular values of $\mathbf{\Sigma}$ for a specific setting where the basis functions ϕ^{ℓ}'s divide the data into disjoint clusters. When this is not the case (e.g. with overlapping clusters (6), or soft clusters (7)), the singular values of $\mathbf{\Sigma}$ would depend on how correlated the basis functions are.

D. Error Bound for FW-EG with Unknown ψ

In this section, we provide an error bound for Algorithm 3* for evaluation metrics of the form $\mathcal{E}^D[h] = \psi(C^D_{11}[h], \ldots, C^D_{mm}[h])$, for a smooth, but *unknown* $\psi : \mathbb{R}^m \to \mathbb{R}_+$. In this case, we do not have a closed-form expression for the gradient of ψ, but instead apply the example weight elicitation routine in Algorithm 1 using probing classifiers chosen from within a small neighborhood around the current iterate h^t, where ψ is effectively linear. Specifically, we invoke Algorithm 1 with the current iterate h^t as the base classifier and with the radius parameter ϵ set to a small value. In the error bound that we state below for this version of the algorithm, we explicitly take into account the "slack" in using a local approximation to ψ as a proxy for its gradient.

Theorem 14 (Error Bound for FW-EG with unknown ψ). *Let $\mathcal{E}^D[h] = \psi(C^D_{11}[h], \ldots, C^D_{mm}[h])$ for an* unknown *concave function $\psi : [0,1]^m \to \mathbb{R}_+$, which is Q-Lipschitz, and also λ-smooth w.r.t. the ℓ_1-norm. Let $\widehat{\mathcal{E}}^{\text{val}}[h] = \psi(\widehat{C}^{\text{val}}_{11}[h], \ldots, \widehat{C}^{\text{val}}_{mm}[h])$. Fix $\delta \in (0,1)$. Suppose Assumption 1 holds with slack ν. Suppose for any linear metric $\sum_i \beta_i C^D_{ii}[h]$, whose associated weight coefficients in the assumption is $\bar{\boldsymbol{\alpha}}$ with $\|\bar{\boldsymbol{\alpha}}\| \leq B$, the following holds. For any $\delta \in (0,1)$, with probability $\geq 1-\delta$ over draw of S^{tr} and S^{val}, when the weight elicitation routine in Algorithm 1 is given an input metric $\widehat{\mathcal{E}}^{\text{val}}$ with $|\widehat{\mathcal{E}}^{\text{val}} - \sum_i \beta_i \widehat{C}^{\text{val}}_{ii}[h]| \leq \chi, \forall h$, it outputs coefficients $\widehat{\boldsymbol{\alpha}}$ such that $\|\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\| \leq \kappa(\delta, n^{\text{tr}}, n^{\text{val}}, \chi)$, for some function $\kappa(\cdot) > 0$. Let $B' = B + \sqrt{Lm}\,\kappa(\delta, n^{\text{tr}}, n^{\text{val}}, 2\lambda\epsilon^2)$. Assume $m \leq n^{\text{val}}$. Then w.p. $\geq 1-\delta$ over draws of S^{tr} and S^{val} from D and μ respectively, the classifier $\widehat{h}$ output by Algorithm 3* with radius parameter ϵ after T iterations satisfies:*

$$\max_h \mathcal{E}^D[h] - \mathcal{E}^D[\widehat{h}] \;\leq\; 2QB'\mathbf{E}_x\left[\|\eta^{\text{tr}}(x) - \widehat{\eta}^{\text{tr}}(x)\|_1\right] + 4Q\sqrt{Lm}\,\kappa(\delta/T, n^{\text{tr}}, n^{\text{val}}, 2\lambda\epsilon^2)$$
$$+\, 4Q\nu \,+\, \mathcal{O}\left(\lambda m\sqrt{\frac{m\log(n^{\text{val}})\log(m) + \log(m/\delta)}{n^{\text{val}}}} + \frac{\lambda}{T}\right).$$

One can plug-in $\kappa(\cdot)$ with e.g. the error bound we derived for Algorithm 1 in Theorem 1, suitably modified to accommodate input metrics $\widehat{\mathcal{E}}^{\text{val}}$ that may differ from the desired linear metric by at most χ. Such modifications can be easily made to Theorem 1 and would result in an additional term $\sqrt{Lm}\chi$ in the error bound to take into account the additional approximation errors in computing the right-hand side of the linear system in (10).

Before proceeding to prove Theorem 14, we state a few useful lemmas. The following lemma shows that because $\psi(\mathbf{C})$ is λ-smooth, it is effectively linear within a small neighborhood around $\mathbf{C}$.

Lemma 15. *Suppose ψ is λ-smooth w.r.t. the ℓ_1-norm. For each iteration t of Algorithm 3*, let $\boldsymbol{v}^t = \nabla\psi(\mathbf{c}^t)$ denote the true gradient of ψ at $\mathbf{c}^t$. Then for any classifier $h^\epsilon(x) = (1-\epsilon)h^t(x) + \epsilon h(x)$,*

$$\left|\widehat{\mathcal{E}}^{\text{val}}[h^\epsilon] - \widehat{\mathcal{E}}^{\text{val}}[h^t] - \sum_i v^t_i \widehat{C}^{\text{val}}_{ii}[h^\epsilon]\right| \;\leq\; 2\lambda\epsilon^2.$$

Proof. For any randomized classifier $h^\epsilon(x) = (1-\epsilon)h^t(x) + \epsilon h(x)$,

$$\begin{aligned}
\left|\widehat{\mathcal{E}}^{\text{val}}[h^\epsilon] - \widehat{\mathcal{E}}^{\text{val}}[h^t] - \sum_i v^t_i \widehat{C}^{\text{val}}_{ii}[h^\epsilon]\right| &= \left|\psi(\text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h^\epsilon])) - \psi(\text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h^t])) - \sum_i v^t_i \widehat{C}^{\text{val}}_{ii}[h^\epsilon]\right| \\
&\leq \frac{\lambda}{2}\|\text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h^\epsilon]) \,-\, \text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h^t])\|_1^2 \\
&= \frac{\lambda}{2}\|\epsilon(\text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h]) \,-\, \text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h^t]))\|_1^2 \\
&= \frac{\lambda}{2}\epsilon^2\|\text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h]) \,-\, \text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h^t])\|_1^2 \\
&\leq \frac{\lambda}{2}\epsilon^2\left(\|\text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h])\|_1 \,+\, \|\text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h^t])\|_1\right)^2 \\
&\leq \frac{\lambda}{2}\epsilon^2(2)^2 \;=\; 2\lambda\epsilon^2.
\end{aligned}$$

Here the second line follows from the fact that ψ is λ-smooth w.r.t. the ℓ_1-norm,and $\boldsymbol{v}^t = \nabla\psi(\text{diag}(\widehat{\mathbf{C}}^{\text{val}}[h^t]))$. The third line follows from linearity of expectations. The last line follows from the fact that the sum of the entries of a confusion matrix (and hence the sum of its diagonal entries) cannot exceed 1. □

We next restate the error bounds for the call to **PI-EW** in line 9 and the corresponding bound on the approximation error in the linear maximizer $\widehat{f}$ obtained.

Lemma 16 (Error bound for call to **PI-EW** in line 9 with unknown ψ)**.** *For each iteration t of Algorithm 3, let $\boldsymbol{\upsilon}^t = \nabla\psi(\mathbf{c}^t)$ denote the true gradient of ψ at $\mathbf{c}^t$, when the algorithm is run with an unknown ψ that is Q-Lipschitz and λ-smooth w.r.t. the ℓ_1-norm. Let $\bar{\boldsymbol{\alpha}}$ be the associated weighting coefficient for the linear metric $\sum_i \upsilon_i^t C_{ii}^D[h]$ (whose coefficients are unknown) in Assumption 1, with $\|\bar{\boldsymbol{\alpha}}\|_1 \leq B$, and with slack ν. Fix $\delta > 0$. Suppose w.p. $\geq 1-\delta$ over draw of S^{tr} and S^{val}, when the weight elicitation routine used in* **PI-EW** *is called with the input metric $\widehat{\mathcal{E}}^{\text{val}}[h] - \widehat{\mathcal{E}}^{\text{val}}[h^t]$ with $|\widehat{\mathcal{E}}^{\text{val}}[h] - \sum_i \upsilon_i \widehat{C}_{ii}^{\text{val}}[h]| \leq \chi, \forall h$, it outputs coefficients $\widehat{\boldsymbol{\alpha}}$ such that $\|\widehat{\boldsymbol{\alpha}} - \bar{\boldsymbol{\alpha}}\| \leq \kappa(\delta, n^{\text{tr}}, n^{\text{val}}, \chi)$, for some function $\kappa(\cdot) > 0$. Let $B' = B + \sqrt{Lm}\,\kappa(\delta, n^{\text{tr}}, n^{\text{val}}, 2\lambda\epsilon^2)$. Then with the same probability, the classifier $\widehat{h}$ output by* **PI-EW** *when called by Algorithm 3* with metric $\widehat{\mathcal{E}}^{\text{lin}}[h] = \widehat{\mathcal{E}}^{\text{val}}[h] - \widehat{\mathcal{E}}^{\text{val}}[h^t]$ and radius ϵ satisfies:*

$$\max_h \sum_i \upsilon_i^t C_{ii}^D[h] - \sum_i \upsilon_i^t C_{ii}^D[\widehat{h}] \quad \leq Q\left(B'\mathbf{E}_x\left[\|\eta^{\text{tr}}(x) - \widehat{\eta}^{\text{tr}}(x)\|_1\right] + 2\sqrt{Lm}\,\kappa(\delta, n^{\text{tr}}, n^{\text{val}}, 2\lambda\epsilon^2) + 2\nu\right),$$

where $\eta_i^{\text{tr}}(x) = \mathbf{P}^\mu(y = i|x)$.

Proof. The proof is the same as that of Lemma 7 for the “known ψ” case, except that the $\kappa(\cdot)$ guarantee for the call to weight elicitation routine in line 2 is different, and takes into account the fact that the input metric $\widehat{\mathcal{E}}^{\text{val}}[h] - \widehat{\mathcal{E}}^{\text{val}}[h^t]$ to the weight elicitation routine is only a local approximation to the (unknown) linear metric $\sum_i \upsilon_i \widehat{C}_{ii}^{\text{val}}[h]$. We use Lemma 15 to compute the value of slack χ in $\kappa(\cdot)$. □

Lemma 17 (Approximation error in linear maximizer $\widehat{f}$ in line 9 with unknown ψ)**.** *For each iteration t in Algorithm 3*, let $\bar{\mathbf{c}}^t = \text{diag}(\mathbf{C}^D[h^t])$ and let $\bar{\boldsymbol{\beta}}^t = \nabla\psi(\bar{\mathbf{c}}^t)$ denote the unknown gradient of ψ evaluated at $\bar{\mathbf{c}}^t$. Suppose the assumptions in Theorem 14 hold. Let $B' = B + \sqrt{Lm}\,\kappa(\delta, n^{\text{tr}}, n^{\text{val}}, 2\lambda\epsilon^2)$. Assume $m \leq n^{\text{val}}$. Then w.p. $\geq 1-\delta$ over draw of S^{tr} and S^{val} from μ and D resp., for any $t = 1, \ldots, T$, the classifier $\widehat{f}$ returned by* **PI-EW** *in line 9 satisfies:*

$$\begin{aligned}\max_h \sum_i \bar{\beta}_{ii}^t C_i^D[h] - \sum_i \bar{\beta}_i^t C_{ii}^D[\widehat{f}] \;\leq\; QB'\mathbf{E}_x\left[\|\eta^{\text{tr}}(x) - \widehat{\eta}^{\text{tr}}(x)\|_1\right] + 2Q\nu \\ + 2Q\sqrt{Lm}\,\kappa\left(\tfrac{\delta}{T}, n^{\text{tr}}, n^{\text{val}}, 2\lambda\epsilon^2\right) + \mathcal{O}\left(\lambda m\sqrt{\frac{m\log(m)\log(n^{\text{val}}) + \log(m/\delta)}{n^{\text{val}}}}\right).\end{aligned}$$

Proof. The proof is the same as that of Lemma 9 for the “known ψ” case, with the only difference being that we use Lemma 16 (instead of Lemma 7) to bound the linear maximization errors in equation (23). □

Proof of Theorem 14. The proof follows from plugging Lemma 17 into the Frank-Wolfe convergence guarantee in Lemma 11 stated in Appendix B.3. □

E. Running Time of Algorithm 3

We discuss how one iteration of FW-EG (Algorithm 3) compares with one iteration (epoch) of training a class-conditional probability estimate $\widehat{\eta}^{\text{tr}}(x) \approx \mathbf{P}^\mu(y = 1|x)$. In each iteration of FW-EG, we create Lm probing classifiers, where each probing classifier via (12) only requires perturbing the predictions of the base classifier $\bar{h} = h^t$ and hence requires $n^{\text{tr}} + n^{\text{val}}$ computations. After constructing the Lm probing classifiers, FW-EG solves a system of linear equations with Lm unknowns, where a naïve matrix inversion approach requires $O((Lm)^3)$ time. Notice that this can be further improved with efficient methods, e.g., using state-of-the-art linear regression solvers. Then FW-EG creates a plugin classifier and combines the predictions with the Frank-Wolfe style updates, requiring $Lm(n^{\text{tr}} + n^{\text{val}})$ computations. So, the overall time complexity for each iteration of FW-EG is $O\left(Lm(n^{\text{tr}} + n^{\text{val}}) + (Lm)^3\right)$. On the other hand, one iteration (epoch) of training $\widehat{\eta}^{\text{tr}}(x)$ requires $O(n^{\text{tr}}Hm)$ time, where H represents the total number of parameters in the underlying model architecture up to the penultimate layer. For deep networks such as ResNets (Sections 7.1 and 7.3), clearly, the run-time is dominated by the training of $\widehat{\eta}^{\text{tr}}(x)$, as long as L and m are relatively small compared to the number of parameters in the neural network.

Thus our approach is reasonably faster than having to train the model for $\widehat{\eta}^{\text{tr}}$ in each iteration (Jiang et al., 2020), training the model (such as ResNets) twice (Patrini et al., 2017), or making multiple forward/backward passes on the training and validation set requiring three times the time for each epoch compared to training $\widehat{\eta}^{\text{tr}}$ (Ren et al., 2018).

F. Plug-in with Coordinate-wise Search Baseline

We describe the Plug-in [train-val] baseline used in Section 7, which constructs a classifier $\widehat{h}(x) \in \operatorname{argmax}_{i\in[m]} w_i\widehat{\eta}_i^{\text{val}}(x)$, by tuning the weights $w_i \in \mathbb{R}$ to maximize the given metric on the validation set . Note that there are m parameters to be tuned, and a naïve approach would be to use an m-dimensional grid search. Instead, we use a trick from Hiranandani et al. (2019b) to decompose this search into an independent coordinate-wise search for each w_i. Specifically, one can estimate the relative weighting w_i/w_j between any pair of classes i, j by constructing a classifier of the form

$$h^{\zeta}(x) = \begin{cases} i & \text{if } \;\zeta\widehat{\eta}_i^{\text{tr}}(x) > (1-\zeta)\widehat{\eta}_j^{\text{tr}}(x) \\ j & \text{otherwise} \end{cases},$$

that predicts either class i or j based on which of these receives a higher (weighted) probability estimates, and (through a line search) finding the parameter $\zeta \in (0, 1)$ for which h^{ζ} yields the highest validation metric:

$$w_i/w_j \approx \operatorname{argmax}_{\zeta\in[0,1]} \widehat{\mathcal{E}}^{\text{val}}[h^{\zeta}].$$

By fixing i to class m, and repeating this for classes $j \in [m-1]$, one can estimate w_j/w_m for each $j \in [m-1]$, and normalize the estimated related weights to get estimates for $w_1, \ldots, w_m$.

G. Solving Constrained Satisfaction Problem in (11)

We describe some common special cases where one can easily identify classifiers $h^{\ell,i}$'s which satisfy the constraints in (11). We will make use of a pre-trained class probability model $\widehat{\eta}_i^{\text{tr}}(x) \approx \mathbf{P}^{\mu}(y = i|x)$, also used in Section 4 to construct the plug-in classifier in Algorithm 2. The hypothesis class $\mathcal{H}$ we consider is the set of all plug-in classifiers obtained by post-shifting $\widehat{\eta}^{\text{tr}}$.

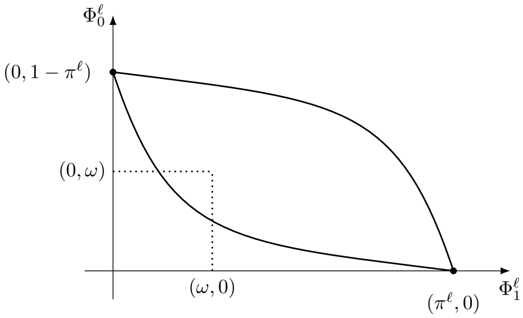

Figure: Geometry of the space of Φ-confusions (Hiranandani et al., 2019a) for $m = 2$ classes and with basis functions $\phi^{\ell}(x) = \mathbf{1}(g(x) = \ell)$ which divide the data into L disjoint clusters. For a fixed cluster ℓ, we plot the values of $\Phi_0^{\mu,\ell}[h]$ and $\Phi_1^{\mu,\ell}[h]$ for all randomized classifiers, with $\pi^{\ell} = \mathbf{P}^{\mu}(y = 1, g(x) = \ell)$. The points on the lower boundary correspond to classifiers of the form $\mathbf{1}(\eta^{\text{tr}}(x) \leq \tau)$ for varying thresholds $\tau \in [0, 1]$. The points on the lower boundary within the dotted box correspond to the thresholded classifiers h which yield both values $\Phi_0^{\mu,\ell}[h] \leq \omega$ and $\Phi_1^{\mu,\ell}[h] \leq \omega$. One can thus find a feasible probing classifier $h^{\ell,i}$ for the constrained optimization problem in (11) using the construction given in (27) as long as $\pi^{\ell} \geq \gamma$ and $1 - \pi^{\ell} \geq \gamma$, and the lower boundary intersects with the dotted box for clusters $\ell' \neq \ell$. If the latter does not happen, one can increase ω in small steps until the classifier given in (27) is feasible for (11).

We start with a binary classification problem ($m = 2$) with basis functions $\phi^{\ell}(x) = \mathbf{1}(g(x) = \ell)$, which divide the data points into L *disjoint* groups according to $g(x) \in [L]$. For this setting, one can show under mild assumptions on the data distribution that (11) does indeed have a feasible solution (using e.g. the geometric techniques used by Hiranandani et al. (2019a) and also elaborated in the figure above). One such feasible $h^{\ell,i}$ predicts class $i \in \{0, 1\}$ on all example belonging to group ℓ, and uses a thresholded of $\widehat{\eta}^{\text{tr}}$ for examples from other groups, with per-cluster thresholds. This would have the effect of maximizing the diagonal entry $\widehat{\Phi}^{\text{tr},\ell}_{i}[h^{\ell,i}]$ of $\widehat{\mathbf{\Sigma}}$ and the thresholds can be tuned so that the off-diagonal entries $\widehat{\Phi}^{\text{tr},\ell'}_{i'}[h^{\ell,i}], \forall(\ell', i') \neq (\ell, i)$ are small. More specifically, for any $\ell \in [L], i \in \{0, 1\}$, the classifier $h^{\ell,i}$ can be constructed as:

$$h^{\ell,i}(x) = \begin{cases} i & \text{if } g(x) = \ell \\ \mathbf{1}(\widehat{\eta}^{\text{tr}}(x) \leq \tau_{g(x)}) & \text{otherwise,} \end{cases} \tag{27}$$

where the thresholds $\tau_{\ell'} \in [0, 1], \ell' \neq \ell$ can each be tuned independently using a line search to minimize $\max_{i'} \widehat{\Phi}^{\text{tr},\ell'}_{i'}[h^{\ell,i}]$. As long as $\widehat{\eta}^{\text{tr}}$ is a close approximation of $\mathbf{P}(y|x)$, the above procedure is guaranteed to find an approximately feasible solution for (11), provided one exists. Indeed one can tune the values of γ and ω in (11), so that the above construction (with tuned thresholds) satisfies the constraints.

We next look a multiclass problem ($m > 2$) with basis functions $\phi^{\ell}(x) = \mathbf{1}(g(x) = \ell)$ which again divide the data points into L *disjoint* groups. Here again, one can show under mild assumptions on the data distribution that (11) does indeed have a feasible solution (using e.g. the geometric tools from Hiranandani et al. (2019b)). We can once again construct a feasible $h^{\ell,i}$ by predicting class $i \in [m]$ on all example belonging to group ℓ, and using a post-shifted classifier for examples from other groups. In particular, for any $\ell \in [L], i \in [m]$, the classifier $h^{\ell,i}$ can be constructed as:

$$h^{\ell,i}(x) = \begin{cases} i & \text{if } g(x) = \ell \\ \operatorname{argmax}_{j \in [m]} w^{g(x)}_{j} \widehat{\eta}^{\text{tr}}_{j}(x) & \text{otherwise} \end{cases}, \tag{28}$$

where we use m parameters $w^{\ell'}_{1}, \ldots, w^{\ell'}_{m}$ for each cluster $\ell' \neq \ell$. We can then tune these m parameters to minimize the maximum of the off-diagonal entries of $\widehat{\mathbf{\Sigma}}$, i.e. minimize $\max_{i'} \widehat{\Phi}^{\text{tr},\ell'}_{i'}[h^{\ell,i}]$. However, this may require an m-dimensional grid search. Fortunately, as described in Appendix F, we can use a trick from Hiranandani et al. (2019b) to reduce the problem of tuning m parameters into m independent line searches. This is based on the idea that the optimal relative weighting $w^{\ell'}_{i}/w^{\ell'}_{j}$ between any pair of classes can be determined through a line search. In our case, we will fix $w^{\ell'}_{m} = 1, \forall \ell' \neq \ell$ and compute $w^{\ell'}_{i}, i = 1, \ldots, m-1$ by solving the following one-dimensional optimization problem to determine the relative weighting $w^{\ell'}_{i}/w^{\ell'}_{m} = w^{\ell'}_{i}$.

$$w^{\ell'}_{i} \in \operatorname*{argmin}_{\zeta \in [0,1]} \left(\max_{i'} \widehat{\Phi}^{\text{tr},\ell'}_{i'}[h^{\zeta}] \right), \quad \text{where} \quad h^{\zeta}(x) = \begin{cases} i & \text{if } \zeta \widehat{\eta}^{\text{tr}}_{i}(x) < (1 - \zeta)\widehat{\eta}^{\text{tr}}_{m}(x) \\ m & \text{otherwise} \end{cases}.$$

We can repeat this for each cluster $\ell' \neq \ell$ to construct the (ℓ, i)-th probing classifier $h^{\ell,i}$ in (28).

For the more general setting, where the basis functions ϕ^{ℓ}'s cluster the data into overlapping or soft clusters (such as in (7)), one can find feasible classifiers for (11) by posing this problem as a "rate" constrained optimization problem of the form below to pick $h^{\ell,i}$:

$$\max_{h \in \mathcal{H}} \widehat{\Phi}^{\text{tr},\ell}_{i}[h] \;\; \text{s.t.} \;\; \widehat{\Phi}^{\text{tr},\ell'}_{i'}[h] \leq \omega, \forall(\ell', i') \neq (\ell, i),$$

which can be solved using off-the-shelf toolboxes such as the open-source library offered by Cotter et al. (2019b).[3] Indeed one can tune the hyper-parameters γ and ω so that the solution to the above problem is feasible for (11). If $\mathcal{H}$ is the set of plug-in classifiers obtained by post-shifting $\widehat{\eta}^{\text{tr}}$, then one can alternatively use the approach of Narasimhan (2018) to identify the optimal post-shift on $\widehat{\eta}^{\text{tr}}$ that solves the above constrained problem.

H. Additional Experimental Details

For the experiments (Section 7), we provide the data statistics in Table 6. Observe that we always use small validation data in comparison to the size of the training data. Below we provide some more details regarding the experiments:

- *Maximizing Accuracy under Label Noise on CIFAR-10 (Section 7.1):* The metric that we aim to optimize is test accuracy, which is a linear metric in the diagonal entries of the confusion matrix. Notice that we work with the *asymmetric* label noise model from Patrini et al. (2017), which corresponds to the setting where a label is flipped to a particular label with a certain probability. This involves a non-diagonal noise transition matrix $\mathbf{T}$, and consequently the

[3] https://github.com/google-research/tensorflow_constrained_optimization

Table 6: Data Statistics for different problem setups in Section 7.

Problem Setup	**Dataset**	**#Classes**	**#Features**	**train / val / test split**
Indepen. Label Noise (Section 7.1)	CIFAR-10	10	$32 \times 32 \times 3$	49K / 1K / 10K
Proxy-Label (Section 7.2)	Adult	2	101	32K / 350 / 16K
Domain-Shift (Section 7.3)	Adience	2	$256 \times 256 \times 3$	12K / 800 / 3K
Black-Box Fairness Metric (Section 7.4)	Adult	2 (2 prot. groups)	106	32K / 1.5K / 14K

corrected training objective is a linear function of the entire confusion matrix. Indeed, the loss correction approach from (Patrini et al., 2017) makes use of the estimate of the entire noise-transition matrix, including the off-diagonal entries. Whereas, our approach in the experiment elicits weights for the diagonal entries alone, but assigns a different set of weights for each basis function, i.e., cluster. We are thus able to achieve better performance than Patrini et al. (2017) by optimizing correcting for the noise using a linear function of per-cluster diagonal entries. Indeed, we also observed that PI-EW often achieves better accuracy during cross-validation with ten basis functions, highlighting the benefit of underlying modeling in PI-EW. We expect to get further improvements by incorporating off-diagonal entries in PI-EW optimization on the training side as explained in Appendix A. We also stress that the results from our methods can be further improved by cross-validating over kernel width, UMAP dimensions, and selection of the cluster centers, which are currently set to fixed values in our experiments. Lastly, we did not compare to the Adaptive Surrogates (Jiang et al., 2020) for this experiment as this baseline requires to re-train the ResNet model in every iteration, and more importantly, this method constructs its probing classifiers by perturbing the parameters of the ResNet model several times in each iteration, which can be prohibitively expensive in practice.

- *Maximizing G-mean with Proxy Labels on Adult (Section 7.2):* In this experiment, we use binary features as basis functions instead of RBF kernels as done in CIFAR-10 experiment. This reflects the flexibility of the proposed PI-EW and FW-EG methods. Our approach can incorporate any indicator features as basis function as long as it reflects cluster memberships. Moreover, our choice of basis function was motivated from choices made in (Jiang et al., 2020). We expect to further improve our results by incorporating more binary features as basis functions.

- *Maximizing F-measure under Domain Shift on Adience (Section 7.3):* As mentioned in Section 7.3, for the basis functions, in addition to the default basis $\phi^{\text{def}}(x) = 1\,\forall x$, we choose from subsets of six basis functions $\phi^1, \ldots, \phi^6$ that are averages of the RBFs, centered at points from the validation set corresponding to each one of the six age-gender combinations. We choose these subsets using knowledge of the underlying image classification task. Specifically, besides the default basis function, we cross-validate over three subsets of basis functions. The first subset comprises two basis functions, where the basis functions are averages of the RBF kernels with cluster centers belonging to the two true class. The second subset comprises three basis functions, where the basis functions are averages of the RBF kernels with cluster centers belonging to the three age-buckets. The third subset comprises six basis functions, where the basis functions are averages of the RBF kernels with cluster centers belonging to the combination of true class × age-bucket. We expect to further improve our results by cross-validating over kernel width and selection of the cluster centers. Lastly, we did not compare to Adaptive Surrogates, as this experiment again requires training a deep neural network model, and perturbing or retraining the model in each iteration can be prohibitively expensive in practice.

- *Maximizing Black-box Fairness Metric on Adult (Section 7.4):* In this experiment, since we treat the metric as a black-box, we do not assume access to gradients and thus do not run the [ψ known] variant of FW-EG. We only report the [ψ unknown] variant of FW-EG with varied basis functions as shown in Table 5.

- In Table 7, we replicate the "Macro F-measure" experiment (without noise) from Section 6.2 in (Jiang et al., 2020) and report results of maximizing the macro F-measure on Adult, COMPAS and Default datasets. We see that our approach yields notable gains on two out of the three datasets in comparison to Adaptive Surrogates approach (Jiang et al., 2020).

Table 7: Test macro F-measure for the maximization task in Section 6.2 of Jiang et al. (2020).

↓ Data, Method →	Adaptive Surrogates (Jiang et al., 2020)	FW-EG
COMPAS	0.629	**0.652**
Adult	0.665	**0.670**
Default	0.533	0.536