A. Related work

Here, we briefly discuss several learning scenarios and work related to our study of federated learning.

The problem of federated learning is closely related to other learning scenarios where there is a mismatch between the source distribution and the target distribution. This includes the problem of *transfer learning* or *domain adaptation* from a single source to a known target domain (Ben-David, Blitzer, Crammer, and Pereira, 2006; Mansour, Mohri, and Rostamizadeh, 2009b; Cortes and Mohri, 2014; Cortes, Mohri, and Muñoz Medina, 2015), either through unsupervised adaptation techniques (Gong et al., 2012; Long et al., 2015; Ganin & Lempitsky, 2015; Tzeng et al., 2015), or via lightly supervised ones (some amount of labeled data from the target domain) (Saenko et al., 2010; Yang et al., 2007; Hoffman et al., 2013; Girshick et al., 2014). This also includes previous applications in natural language processing (Dredze et al., 2007; Blitzer et al., 2007; Jiang & Zhai, 2007; Raju et al., 2018), speech recognition (Legetter & Woodland, 1995; Gauvain & Chin-Hui, 1994; Pietra et al., 1992; Rosenfeld, 1996; Jelinek, 1998; Roark & Bacchiani, 2003), and computer vision (Martínez, 2002)

A problem more closely related to that of federated learning is that of *multiple-source adaptation*, first formalized and analyzed theoretically by Mansour, Mohri, and Rostamizadeh (2009c;a) and later studied for various applications such as object recognition (Hoffman et al., 2012; Gong et al., 2013a;b). Recently, (Zhang et al., 2015) studied a causal formulation of this problem for a classification scenario, using the same combination rules as Mansour et al. (2009c;a). The problem of *domain generalization* (Pan & Yang, 2010; Muandet et al., 2013; Xu et al., 2014), where knowledge from an arbitrary number of related domains is combined to perform well on a previously unseen domain is very closely related to that of federated learning, though the assumptions about the information available to the learner and the availability of unlabeled data may differ.

In the multiple-source adaptation problem studied by Mansour, Mohri, and Rostamizadeh (2009c;a) and Hoffman, Mohri, and Zhang (2018), each domain k is defined by the corresponding distribution $\mathcal{D}_k$ and the learner has only access to a predictor h_k for each domain and no access to labeled training data drawn from these domains. The authors show that it is possible to define a predictor h whose expected loss $\mathcal{L}_{\mathcal{D}}(h)$ with respect to any distribution $\mathcal{D}$ that is a mixture of the source domains $\mathcal{D}_k$ is at most the maximum expected loss of the source predictors: $\max_k L_{\mathcal{D}_k}(h_{\mathcal{D}_k})$. They also provide an algorithm for determining h.

Our learning scenario differs from the one adopted in that work since we assume access to labeled training data from each domain $\mathcal{D}_k$. Furthermore, the predictor determined by the algorithm of Hoffman, Mohri, and Zhang (2018) belongs to a specific hypothesis set $\mathcal{H}'$, which is that of distribution weighted combinations of the domain predictors h_k, while, in our setup, the objective is to determine the best predictor in some global hypothesis set $\mathcal{H}$, which may include $\mathcal{H}'$ as a subset, and which is not depending on some domain-specific predictors.

Our optimization solution also differs from the work of Farnia & Tse (2016) and Lee & Raginsky (2017) on local minimax results, where samples are drawn from a single source $\mathcal{D}$, and where the generalization error is minimized over a set of locally ambiguous distributions $\widehat{\mathcal{D}}$, where $\widehat{\mathcal{D}}$ is the empirical distribution. The authors propose this metric for statistical robustness. In our work, we obtain samples from p unknown distributions, and the set of distributions D_λ over which we optimize the expected loss is fixed and independent of samples. Furthermore, the source distributions can differ arbitrarily and need not be close to each other. In reverse, we note that our stochastic algorithm can be used to minimize the loss functions proposed in (Farnia & Tse, 2016; Lee & Raginsky, 2017).

B. Extensions

In this section, we briefly discuss several extensions of the framework, theory and algorithms that we presented.

B.1. Domain definitions

The choice of the domains can significantly impact learnability in federated learning. In view of our learning bounds, if the number of domains, p, is large and Λ is the full simplex, $\Lambda = \Delta_p$, then the models may not generalize well. Thus, if the number of clients is very large, using each client as a domain may be a poor choice for better generalization. Ideally, each domain is represented with a sufficiently large number of samples and is relatively homogeneous or pure. This suggests using a clustering algorithm for defining the domains based on the similarity of the client distributions. Different Bregman divergences could be used to define the divergence or similarity between distributions. Thus, techniques such as those of

Banerjee, Merugu, Dhillon, and Ghosh (2005) could be used to determine clusters of clients using a suitable Bregman divergence.

Client clusters can also be determined based on domain expertise. For example, in federated keyboard next word prediction (Hard et al., 2018), domains can be chosen to be the native language of the clients. If the model is used in variety of applications, domains can also be based on the application of interest. For example, the keyboard in (Hard et al., 2018) is used in chat apps, long form text input apps, and web inputs. Here, domains can be the app that was used. Training models agnostically ensures that the user experience is favorable in all apps.

B.2. Incorporating a prior on Λ

Agnostic federated learning as defined in (1) treats all domains equally and does not incorporate any prior knowledge of λ. Suppose we have a prior distribution $p_\Lambda(\lambda)$ over $\lambda \in \Lambda$ at our disposal, then, we can modify (1) to incorporate that prior. If the loss function ℓ is the cross-entropy loss, then the agnostic loss can be modified as follows:

$$\max_{\lambda \in \Lambda} \left(\mathcal{L}_{D_\lambda}(h) + \log p_\Lambda(\lambda) \right). \tag{9}$$

In this formulation, larger weights are assigned to more likely domains. The generalization guarantees of Theorem 1 can be appropriately modified to include these changes. Furthermore, if the prior $p_\Lambda(\lambda)$ is a log-concave function of λ, then the new objective is convex in h and concave in λ and a slight modification of our proposed algorithm can be used to determine the global minima. We note that we could also adopt a multiplicative formulation with the prior multiplying the loss, instead of the additive one with the negative log of the probability in Equation 9.

B.3. Domain features and personalization

We studied agnostic federated learning, where we learn a model that performs well on all domains. First, notice that we do not make any assumption on the hypothesis set $\mathcal{H}$ and the hypotheses can use the domain k as a feature. Such models could be useful for applications where the target domain is known at inference time. Second, while this paper deals with learning a centralized model, the resulting model $h_{\mathcal{D}_\Lambda}$ can be combined with a personalized model, on the client's machine, to design better client-specific models. This can be done for example by learning an appropriate mixture weight $\alpha_k \in [0,1]$ to use a mixture $\alpha_k h_{\mathcal{D}_\Lambda} + (1-\alpha_k) h_k$ of the domain agnostic centralized model $h_{\mathcal{D}_\Lambda}$ and a client- or domain-specific model h_k.

C. Learning-theoretical guarantees

C.1. Proof of Proposition 1

Consider the following two distributions with support reduced to a single element $x \in \mathcal{X}$ and two classes $\mathcal{Y} = \{0,1\}$: $\mathcal{D}_1(x,0) = 0$, $\mathcal{D}_1(x,1) = 1$, $\mathcal{D}_2(x,0) = \frac{1}{2}$, and $\mathcal{D}_2(x,1) = \frac{1}{2}$. Let $\Lambda = \{\delta_1, \delta_2\}$, where δ_k, $k = 1,2$, denotes the Dirac measure on index k. We will consider the case where the sample sizes m_k are all equal, that is $h_{\overline{\mathcal{U}}} = \frac{1}{2}(\mathcal{D}_1 + \mathcal{D}_2)$. Let p_0 denote the probability that h assigns to class 0 and p_1 the one it assigns to class 1. Then, the cross-entropy loss of a predictor h can be expressed as follows:

$$\begin{aligned}
\mathcal{L}_{\overline{\mathcal{U}}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\overline{\mathcal{U}}}\big[-\log p_y\big] &= \frac{1}{4}\log\frac{1}{p_0} + \frac{1}{2}\log\frac{1}{p_1} + \frac{1}{4}\log\frac{1}{p_1} \\
&= \frac{1}{4}\log\frac{1}{p_0} + \frac{3}{4}\log\frac{1}{p_1} \\
&= \mathsf{D}\big(\big(\tfrac{1}{4},\tfrac{3}{4}\big) \parallel (p_0,p_1)\big) + \frac{1}{4}\log\frac{4}{1} + \frac{3}{4}\log\frac{4}{3} \\
&\geq \frac{1}{4}\log\frac{4}{1} + \frac{3}{4}\log\frac{4}{3},
\end{aligned}$$

where the last inequality follows the non-negativity of the relative entropy. Furthermore, equality is achieved when $p_0 = 1 - p_1 = \frac{1}{4}$, which defines $h_{\overline{\mathcal{U}}}$, the minimizer of $\mathcal{L}_{\overline{\mathcal{U}}}(h)$. In view of that, $\mathcal{L}_{\mathcal{D}_\Lambda}(h_{\overline{\mathcal{U}}})$ is given by the following:

$$\begin{aligned}
\mathcal{L}_{\mathcal{D}_\Lambda}(h_{\overline{\mathcal{U}}}) &= \max\left(\mathcal{L}_{\delta_1}(\overline{\mathcal{U}}), \mathcal{L}_{\delta_2}(\overline{\mathcal{U}})\right)\\
&= \max\left\{\log\frac{4}{3}, \frac{1}{2}\log\frac{4}{1} + \frac{1}{2}\log\frac{4}{3}\right\}\\
&= \log\frac{4}{\sqrt{3}}.
\end{aligned}$$

We now compute the loss of $h_{\mathcal{D}_\Lambda}$:

$$\begin{aligned}
\min_{h\in\mathcal{H}}\mathcal{L}_{\mathcal{D}_\Lambda}(h) &= \min_{h\in\mathcal{H}}\max_{k\in[p]}\mathcal{L}_{\mathcal{D}_k}(h)\\
&= \min_{(p_0,p_1)\in\Delta_2}\max\left\{\log\frac{1}{p_1}, \frac{1}{2}\log\frac{1}{p_0} + \frac{1}{2}\log\frac{1}{p_1}\right\}\\
&= \min_{p_1\in[0,1]}\max\left\{\log\frac{1}{p_1}, \log\frac{1}{\sqrt{p_1(1-p_1)}}\right\}\\
&= \log 2,
\end{aligned}$$

since $\frac{1}{2}$ is the solution of the convex optimization in p_1, in view of $\max\left\{\frac{1}{p_1}, \frac{1}{\sqrt{p_1(1-p_1)}}\right\} = \frac{1}{\sqrt{p_1(1-p_1)}} \leq \frac{1}{2}$ for $p_1 > \frac{1}{2}$.

C.2. Proof of Theorem 1

The proof is an extension of the standard proofs for Rademacher complexity generalization bounds (Koltchinskii & Panchenko, 2002; Mohri et al., 2018). Fix $\lambda \in \Lambda$. For any sample $S = S_1, \ldots, S_p$, define $\Psi(S_1, \ldots, S_p)$ by

$$\Psi(S_1, \ldots, S_p) = \sup_{h\in\mathcal{H}}\left(\mathcal{L}_{\mathcal{D}_\lambda}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)\right).$$

Let $S' = (S'_1, \ldots, S'_p)$ be a sample differing from $S = (S_1, \ldots, S_p)$ only by point $x'_{k,i}$ in S'_k and $x_{k,i}$ in S_k. Then, since the difference of suprema over the same set is bounded by the supremum of the differences, we can write

$$\begin{aligned}
\Psi(S') - \Psi(S) &= \sup_{h\in\mathcal{H}}\left(\mathcal{L}_{\mathcal{D}_\lambda}(h) - \mathcal{L}_{\overline{\mathcal{D}}'_\lambda}(h)\right) - \sup_{h\in\mathcal{H}}\left(\mathcal{L}_{\mathcal{D}_\lambda}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)\right)\\
&\leq \sup_{h\in\mathcal{H}}\left(\mathcal{L}_{\mathcal{D}_\lambda}(h) - \mathcal{L}_{\overline{\mathcal{D}}'_\lambda}(h)\right) - \left(\mathcal{L}_{\mathcal{D}_\lambda}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)\right)\\
&\leq \sup_{h\in\mathcal{H}}\mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h) - \mathcal{L}_{\overline{\mathcal{D}}'_\lambda}(h)\\
&= \sup_{h\in\mathcal{H}}\sum_{k=1}^{p}\frac{\lambda_k}{m_k}\sum_{i=1}^{m_k}\ell(h(x'_{k,i}), y'_{k,i}) - \sum_{k=1}^{p}\frac{\lambda_k}{m_k}\sum_{i=1}^{m_k}\ell(h(x_{k,i}), y_{k,i})\\
&= \sup_{h\in\mathcal{H}}\frac{\lambda_k}{m_k}\left[\ell(h(x'_{k,i}), y'_{k,i}) - \ell(h(x_{k,i}), y_{k,i})\right]\\
&\leq \frac{\lambda_k M}{m_k}.
\end{aligned}$$

Thus, by McDiarmid's inequality, for any $\delta > 0$, the following inequality holds with probability at least $1 - \delta$ for any $h \in \mathcal{H}$:

$$\mathcal{L}_{\mathcal{D}_\lambda}(h) \leq \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h) + \mathbb{E}\left[\max_{h\in\mathcal{H}}\mathcal{L}_{\mathcal{D}_\lambda}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)\right] + M\sqrt{\sum_{k=1}^{p}\frac{\lambda_k^2}{2m_k}\log\frac{1}{\delta}}.$$

Therefore, by the union over Λ_ϵ, with probability at least $1 - \delta$, for any $h \in \mathcal{H}$ and $\lambda \in \Lambda_\epsilon$ the following holds:

$$\mathcal{L}_{\mathcal{D}_\lambda}(h) \leq \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h) + \mathbb{E}\left[\max_{h\in\mathcal{H}}\mathcal{L}_{\mathcal{D}_\lambda}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)\right] + M\sqrt{\sum_{k=1}^{p}\frac{\lambda_k^2}{2m_k}\log\frac{|\Lambda_\epsilon|}{\delta}}.$$

By definition of Λ_ϵ, for any $\lambda \in \Lambda$, there exists $\lambda' \in \Lambda_\epsilon$ such that $\mathcal{L}_{\mathcal{D}_\lambda}(h) \leq \mathcal{L}_{\mathcal{D}'_\lambda}(h) + M\epsilon$. In view of that, with probability at least $1 - \delta$, for any $h \in \mathcal{H}$ and $\lambda \in \Lambda$ the following holds:

$$\mathcal{L}_{\mathcal{D}_\lambda}(h) \leq \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h) + \mathbb{E}\left[\max_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}_\lambda}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)\right] + M\epsilon + M\sqrt{\sum_{k=1}^{p} \frac{\lambda_k^2}{2m_k} \log \frac{|\Lambda_\epsilon|}{\delta}}.$$

The expectation appearing on the right-hand side can be bounded following standard proofs for Rademacher complexity upper bounds (see for example (Mohri et al., 2018)), leading to

$$\mathbb{E}\left[\max_{h \in \mathcal{H}} \mathcal{L}_{\mathcal{D}_\lambda}(h) - \mathcal{L}_{\overline{\mathcal{D}}_\lambda}(h)\right] \leq \mathfrak{R}_{\mathbf{m}}(\mathcal{G}, \lambda).$$

The sum $\sum_{k=1}^{p} \frac{\lambda_k^2}{m_k}$ can be expressed in terms of the skewness of Λ, using the following equalities:

$$m \sum_{k=1}^{p} \frac{\lambda_k^2}{m_k} = \sum_{k=1}^{p} \frac{\lambda_k^2}{\frac{m_k}{m}} = \sum_{k=1}^{p} \frac{\lambda_k^2}{\frac{m_k}{m}} + \sum_{k=1}^{p} \frac{m_k}{m} - 2\sum_{k=1}^{p} \lambda_k + 1 = \sum_{k=1}^{p} \frac{(\lambda_k - \frac{m_k}{m})^2}{\frac{m_k}{m}} + 1 = \chi^2(\lambda \,\|\, \overline{\mathbf{m}}) + 1.$$

This completes the proof.

C.3. Proof of Lemma 1

For any $\lambda \in \Lambda$, define the set of vectors A_λ in $\mathbb{R}^m$ by

$$A_\lambda = \left\{ \left[\frac{\lambda_k}{m_k} \ell(h(x_{k,i}), y_{k,i}) \right]_{(k,i) \in [p] \times [m_k]} : \mathbf{x} \in \mathcal{X}^m, \mathbf{y} \in \mathcal{Y}^m \right\}.$$

For any $\mathbf{a} \in A_\lambda$, $\|\mathbf{a}\|_2 = \sqrt{\sum_{k=1}^{p} m_k \frac{\lambda_k^2}{m_k^2}} = \sqrt{\sum_{k=1}^{p} \frac{\lambda_k^2}{m_k}} \leq \sqrt{\frac{\mathfrak{s}(\Lambda \,\|\, \overline{\mathbf{m}})}{m}}$. Then, by Massart's lemma, for any $\lambda \in \Lambda$, the following inequalities hold:

$$\begin{aligned}
\mathfrak{R}_{\mathbf{m}}(\mathcal{G}, \lambda) &= \mathop{\mathbb{E}}_{\substack{S_k \sim \mathcal{D}_k^{m_k} \\ \boldsymbol{\sigma}}} \left[\sup_{h \in \mathcal{H}} \sum_{k=1}^{p} \frac{\lambda_k}{m_k} \sum_{i=1}^{m_k} \sigma_{k,i}\, \ell(h(x_{k,i}), y_{k,i}) \right] \\
&\leq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\mathbf{a} \in A} \sum_{k=1}^{p} \sum_{i=1}^{m_k} \sigma_{k,i} a_{k,i} \right] \\
&\leq \sqrt{\frac{\mathfrak{s}(\Lambda \,\|\, \overline{\mathbf{m}})}{m}} \frac{\sqrt{2 \log |A_\lambda|}}{m} \\
&= \frac{\sqrt{2\mathfrak{s}(\Lambda \,\|\, \overline{\mathbf{m}}) \log |A_\lambda|}}{m}.
\end{aligned}$$

By Sauer's lemma, the following holds for $m \geq d$: $|A_\lambda| \leq \left(\frac{em}{d}\right)^d$. Plugging in the right-hand side in the inequality above completes the proof.

D. Alternative learning guarantees

An objective similar to that of AFL was considered in the context of multiple-source domain adaptation by Liu et al. (2015). The authors presented generalization bounds for a scenario where the target is based on some specific mixture λ of the source domains. Our theoretical results differ from those of this work in two ways. First, our generalization bounds do not hold for a single mixture weight λ but for any subset Λ of the simplex. Second, the complexity terms in the bounds presented by these authors are proportional to $\sqrt{m} \max_{k \in [p]} \frac{\lambda_k}{m_k}$, while our guarantees are in terms of $\sqrt{\sum_{k=1}^{p} \frac{\lambda_k^2}{m_k}}$, which is strictly tighter. In particular, in the special case where $k = 2$, $\lambda_1 = \frac{1}{\sqrt{m}}$, $\lambda_2 = 1 - \lambda_1$ and $m_1 = 1$ and $m_2 = m - 1$, the bounds of Liu et al. (2015) are proportional to a constant and thus not informative, $\sqrt{m} \max_{k \in [p]} \frac{\lambda_k}{m_k} = 1$, while our guarantees are in terms of $\frac{1}{\sqrt{m}}$.

Our generalization error in Theorem 1 is particularly useful when Λ is a strict subset of the simplex, $\Lambda \subset \Delta_p$. If $\Lambda = \Delta_p$, we can give the following alternative learning guarantee based.

Theorem 3. *For any $\delta > 0$, with probability at least $1 - \delta$ over the draw of samples $S_k \sim \mathcal{D}_k^{m_k}$, the following inequality holds for all $h \in \mathcal{H}$ and $\lambda \in \Lambda$:*

$$L_{\mathcal{D}_\lambda}(h) \le L_{\overline{\mathcal{D}}_\lambda}(h) + \sum_{k=1}^{p} \left(2\lambda_k \mathfrak{R}^k_{m_k}(\mathcal{G}) + \lambda_k M \sqrt{\frac{1}{2m_k} \log \frac{p}{\delta}} \right),$$

where $\mathfrak{R}^k_{m_k}(\mathcal{G})$ is the Rademacher complexity over domain $\mathcal{D}_k$ with m_k samples.

Proof. For a fixed $k \in [p]$, by a standard Rademacher complexity bound, for any $\delta > 0$, with probability at least $1 - \delta$, the following inequality holds for all $h \in \mathcal{H}$:

$$L_{\mathcal{D}_k}(h) \le L_{\overline{\mathcal{D}}_k}(h) + 2\mathfrak{R}^k_{m_k}(\mathcal{G}) + M\sqrt{\frac{1}{2m_k} \log \frac{1}{\delta}}.$$

Summing up the inequalities for each $k \in [p]$ after multiplication by λ_k and using the union bound complete the proof. □

We will now compare the generalization bounds of Theorem 1 and Theorem 3. The Rademacher complexity term of the bound of Theorem 1, $\mathfrak{R}_{\mathbf{m}}(\mathcal{G}, \lambda)$, is more favorable than that of Theorem 3, since by the sub-additivity of sup and the linearity of expectation, we can write

$$\begin{aligned}
\mathfrak{R}_{\mathbf{m}}(\mathcal{G}, \lambda) = \mathop{\mathbb{E}}_{\substack{S_k \sim \mathcal{D}_k^{m_k} \\ \boldsymbol{\sigma}}} \left[\sup_{h \in \mathcal{H}} \sum_{k=1}^{p} \frac{\lambda_k}{m_k} \sum_{i=1}^{m_k} \sigma_{k,i}\, \ell(h(x_{k,i}), y_{k,i}) \right] &\le \mathop{\mathbb{E}}_{\substack{S_k \sim \mathcal{D}_k^{m_k} \\ \boldsymbol{\sigma}}} \left[\sum_{k=1}^{p} \frac{\lambda_k}{m_k} \sup_{h \in \mathcal{H}} \sum_{i=1}^{m_k} \sigma_{k,i}\, \ell(h(x_{k,i}), y_{k,i}) \right] \\
&= \sum_{k=1}^{p} \frac{\lambda_k}{m_k} \mathop{\mathbb{E}}_{\substack{S_k \sim \mathcal{D}_k^{m_k} \\ \boldsymbol{\sigma}}} \left[\sup_{h \in \mathcal{H}} \sum_{i=1}^{m_k} \sigma_{k,i}\, \ell(h(x_{k,i}), y_{k,i}) \right] \\
&= \sum_{k=1}^{p} \lambda_k \mathfrak{R}^k_{m_k}(\mathcal{G}).
\end{aligned}$$

The comparison of the last terms of the two bounds, $M\sqrt{\frac{\mathfrak{s}(\lambda \,\|\, \overline{\mathbf{m}})}{2m} \log \frac{|\Lambda_\epsilon|}{\delta}}$ versus $M \sum_{k=1}^{p} \sqrt{\frac{1}{2m_k} \log \frac{p}{\delta}}$, depends on the covering number $|\Lambda_\epsilon|$. When $|\Lambda_\epsilon|$ is small, as in the case where Λ is a finite discrete set (in the extreme case reduced to a single element), then, the last term of the bound of Theorem 1 is more favorable. This is because $|\Lambda_\epsilon|$ is then smaller or in the same order of magnitude as p, while, by the sub-additivity of $\sqrt{\cdot}$, the following inequality holds:

$$\sqrt{\frac{\mathfrak{s}(\lambda \,\|\, \overline{\mathbf{m}})}{m}} = \sqrt{\sum_{k=1}^{p} \frac{\lambda_k^2}{m_k}} \le \sum_{k=1}^{p} \sqrt{\frac{\lambda_k^2}{m_k}} = \sum_{k=1}^{p} \lambda_k \sqrt{\frac{1}{m_k}}.$$

On the other hand, when $|\Lambda_\epsilon| = O((\frac{1}{\epsilon})^p)$ as in the case where Λ is the full simplex, then $\log |\Lambda_\epsilon| = p\, O(\log \frac{1}{\epsilon})$ can be substantially larger than $\log p$ and the last term of the bound of Theorem 3 seems more favorable since, by the Cauchy-Schwarz inequality, the following inequality holds:

$$\sum_{k=1}^{p} \lambda_k \sqrt{\frac{1}{m_k}} \le \sqrt{p} \sqrt{\sum_{k=1}^{p} \frac{\lambda_k^2}{m_k}} = \sqrt{p} \sqrt{\frac{\mathfrak{s}(\lambda \,\|\, \overline{\mathbf{m}})}{m}}.$$

In general, it is not clear which of the two bounds is more favorable. This depends on $\overline{\mathbf{m}}$, λ, and Λ_ϵ. Learning bounds improving upon both may be based on a careful interpolation, which we leave to future work.

E. Analysis of the optimization algorithm

E.1. Proof of Theorem 4

The time complexity of the algorithm follows the definitions of the complexity terms U_λ, U_w, and U_p the dimension d in Properties 1. To prove the convergence guarantee, we first state the following lemma.

Lemma 5. *Assume that the Property 1.1 holds. Then,*

$$\max_{\lambda \in \Lambda} \mathsf{L}(w^A, \lambda) - \min_{w \in \mathcal{W}} \max_{\lambda \in \Lambda} \mathsf{L}(w, \lambda) \le \frac{1}{T} \max_{\substack{\lambda \in \Lambda \\ w \in \mathcal{W}}} \left\{ \sum_{t=1}^{T} \mathsf{L}(w_t, \lambda) - \mathsf{L}(w, \lambda_t) \right\}.$$

Proof. Recall that (w^A, λ^A) is the solution returned by the algorithm. First observe that L is convex in w and linear and thus concave in λ. Thus, by the generalized von Neumann's theorem, the following holds:

$$\begin{aligned}
\max_{\lambda \in \Lambda} \mathsf{L}(w^A, \lambda) - \min_{w \in \mathcal{W}} \max_{\lambda \in \Lambda} \mathsf{L}(w, \lambda) &= \max_{\lambda \in \Lambda} \mathsf{L}(w^A, \lambda) - \max_{\lambda \in \Lambda} \min_{w \in \mathcal{W}} \mathsf{L}(w, \lambda) && \text{(von Neumann's minimax)} \\
&\le \max_{\lambda \in \Lambda} \left\{ \mathsf{L}(w^A, \lambda) - \min_{w \in \mathcal{W}} \mathsf{L}(w, \lambda^A) \right\} && \text{(subadd. of } \max\text{)} \\
&= \max_{\substack{\lambda \in \Lambda \\ w \in \mathcal{W}}} \left\{ \mathsf{L}(w^A, \lambda) - \mathsf{L}(w, \lambda^A) \right\} \\
&\le \frac{1}{T} \max_{\substack{\lambda \in \Lambda \\ w \in \mathcal{W}}} \left\{ \sum_{t=1}^{T} \mathsf{L}(w_t, \lambda) - \mathsf{L}(w, \lambda_t) \right\}. && \text{(convexity in } w \text{ and lin. in } \lambda\text{)}
\end{aligned}$$

This completes the proof. □

In view of the lemma, to derive convergence guarantees for the algorithm, it suffices to bound $\mathsf{L}(w_t, \lambda) - \mathsf{L}(w, \lambda_t)$. Since L is linear in λ and convex in w, we have

$$\begin{aligned}
\mathsf{L}(w_t, \lambda) - \mathsf{L}(w, \lambda_t) &= \mathsf{L}(w_t, \lambda) - \mathsf{L}(w_t, \lambda_t) + \mathsf{L}(w_t, \lambda_t) - \mathsf{L}(w, \lambda_t) \\
&\le (\lambda - \lambda_t) \nabla_\lambda \mathsf{L}(w_t, \lambda_t) + (w_t - w) \nabla_w \mathsf{L}(w_t, \lambda_t) \\
&\le (\lambda - \lambda_t) \delta_\lambda \mathsf{L}(w_t, \lambda_t) + (w_t - w) \delta_w \mathsf{L}(w_t, \lambda_t) \\
&+ (\lambda - \lambda_t)(\nabla_\lambda \mathsf{L}(w_t, \lambda_t) - \delta_\lambda \mathsf{L}(w_t, \lambda_t)) + (w_t - w)(\nabla_w \mathsf{L}(w_t, \lambda_t) - \delta_w \mathsf{L}(w_t, \lambda_t)).
\end{aligned}$$

In view of these inequalities, by the subadditivity of $\max$, the following inequality holds:

$$\begin{aligned}
&\max_{\substack{\lambda \in \Lambda \\ w \in \mathcal{W}}} \left\{ \sum_{t=1}^{T} \mathsf{L}(w_t, \lambda) - \mathsf{L}(w, \lambda_t) \right\} \\
&\le \max_{\substack{\lambda \in \Lambda \\ w \in \mathcal{W}}} \sum_{t=1}^{T} (\lambda - \lambda_t) \delta_\lambda \mathsf{L}(w_t, \lambda_t) + (w_t - w) \delta_w \mathsf{L}(w_t, \lambda_t) \\
&+ \max_{\substack{\lambda \in \Lambda \\ w \in \mathcal{W}}} \sum_{t=1}^{T} \lambda (\nabla_\lambda \mathsf{L}(w_t, \lambda_t) - \delta_\lambda \mathsf{L}(w_t, \lambda_t)) - w (\nabla_w \mathsf{L}(w_t, \lambda_t) - \delta_w \mathsf{L}(w_t, \lambda_t)) \\
&+ \sum_{t=1}^{T} \lambda_t (\nabla_\lambda \mathsf{L}(w_t, \lambda_t) - \delta_\lambda \mathsf{L}(w_t, \lambda_t)) - w_t (\nabla_w \mathsf{L}(w_t, \lambda_t) - \delta_w \mathsf{L}(w_t, \lambda_t)).
\end{aligned} \tag{10}$$

We now bound each of the three terms above separately. For the first term, observe that for any $w \in \mathcal{W}$,

$$\begin{aligned}
&\sum_{t=1}^{T}(w_t - w)\delta_w \mathsf{L}(w_t, \lambda_t) \\
&= \frac{1}{2\gamma_w}\sum_{t=1}^{T}\|(w_t - w)\|_2^2 + \gamma_w^2\|\delta_w \mathsf{L}(w_t, \lambda_t)\|_2^2 - \|(w_t - \gamma_w \delta_w \mathsf{L}(w_t, \lambda_t) - w)\|_2^2 \\
&\leq \frac{1}{2\gamma_w}\sum_{t=1}^{T}\|(w_t - w)\|_2^2 + \gamma_w^2\|\delta_w \mathsf{L}(w_t, \lambda_t)\|_2^2 - \|(w_{t+1} - w)\|_2^2 && \text{(property of projection)} \\
&= \frac{1}{2\gamma_w}\|(w_1 - w)\|_2^2 - \|(w_{T+1} - w)\|_2^2 + \frac{\gamma_w}{2}\sum_{t=1}^{T}\|\delta_w \mathsf{L}(w_t, \lambda_t)\|_2^2 && \text{(telescoping sum)} \\
&\leq \frac{1}{2\gamma_w}\|(w_1 - w)\|_2^2 + \frac{\gamma_w}{2}\sum_{t=1}^{T}\|\delta_w \mathsf{L}(w_t, \lambda_t)\|_2^2 \\
&\leq \frac{2R_{\mathcal{W}}^2}{\gamma_w} + \frac{\gamma_w}{2}\sum_{t=1}^{T}\|\delta_w \mathsf{L}(w_t, \lambda_t)\|_2^2 \\
&\leq \frac{2R_{\mathcal{W}}^2}{\gamma_w} + \frac{\gamma_w}{2}\sum_{t=1}^{T}\|\delta_w \mathsf{L}(w_t, \lambda_t) - \nabla_w \mathsf{L}(w_t, \lambda_t) + \nabla_w \mathsf{L}(w_t, \lambda_t)\|_2^2.
\end{aligned}$$

Since the right-hand side does not depend on w, taking the maximum of both sides over $w \in \mathcal{W}$ and the expectation yields

$$\mathbb{E}\left[\max_{w\in\mathcal{W}}\sum_{t=1}^{T}(w_t - w)\delta_w \mathsf{L}(w_t, \lambda_t)\right] \leq \frac{2R_{\mathcal{W}}^2}{\gamma_w} + \frac{\gamma_w T\sigma_w^2}{2} + \frac{T\gamma_w G_w^2}{2},$$

using the following identity:

$$\begin{aligned}
&\mathbb{E}\left[\|\delta_w \mathsf{L}(w_t, \lambda_t) - \nabla_w \mathsf{L}(w_t, \lambda_t) + \nabla_w \mathsf{L}(w_t, \lambda_t)\|_2^2\right] \\
&= \mathbb{E}\left[\|\delta_w \mathsf{L}(w_t, \lambda_t) - \nabla_w \mathsf{L}(w_t, \lambda_t)\|^2\right] - 2\,\mathbb{E}\left[\delta_w \mathsf{L}(w_t, \lambda_t) - \nabla_w \mathsf{L}(w_t, \lambda_t)\right]\cdot\nabla_w \mathsf{L}(w_t, \lambda_t) + \|\nabla_w \mathsf{L}(w_t, \lambda_t)\|_2^2 \\
&= \mathbb{E}\left[\|\delta_w \mathsf{L}(w_t, \lambda_t) - \nabla_w \mathsf{L}(w_t, \lambda_t)\|^2\right] + \|\nabla_w \mathsf{L}(w_t, \lambda_t)\|_2^2.
\end{aligned}$$

Similarly, using the projection property, the following inequality can be shown:

$$\mathbb{E}\left[\max_{\lambda\in\Lambda}\sum_{t=1}^{T}(\lambda - \lambda_t)\delta_\lambda \mathsf{L}(w_t, \lambda_t)\right] \leq \frac{2R_{\Lambda}^2}{\gamma_\lambda} + \frac{\gamma_\lambda T\sigma_\lambda^2}{2} + \frac{T\gamma_\lambda G_\lambda^2}{2}.$$

For the second term, by the Cauchy-Schwarz inequality, we can write

$$\max_{\lambda\in\Lambda}\sum_{t=1}^{T}\lambda(\nabla_\lambda \mathsf{L}(w_t, \lambda_t) - \delta_\lambda \mathsf{L}(w_t, \lambda_t)) \leq R_\Lambda\|\sum_{t=1}^{T}\nabla_\lambda \mathsf{L}(w_t, \lambda_t) - \delta_\lambda \mathsf{L}(w_t, \lambda_t)\|_2.$$

Taking the expectation of both sides and using Jensen's inequality yields

$$\mathbb{E}\left[\max_{\lambda\in\Lambda}\sum_{t=1}^{T}\lambda(\nabla_\lambda \mathsf{L}(w_t, \lambda_t) - \delta_\lambda \mathsf{L}(w_t, \lambda_t))\right] \leq R_\Lambda\sqrt{T}\sigma_\lambda.$$

Similarly, we obtain the following:

$$\mathbb{E}\left[\max_{w\in\mathcal{W}} w\nabla_w \mathsf{L}(w_t, \lambda_t) - \delta_w \mathsf{L}(w_t, \lambda_t)\right] \leq R_{\mathcal{W}}\sqrt{T}\sigma_w.$$

For the third term, observe that the stochastic gradients at time t are unbiased, conditioned on λ_t, and w_t, hence,

$$\mathbb{E}\left[\sum_{t=1}^{T}\lambda_t(\nabla_\lambda \mathsf{L}(w_t, \lambda_t) - \delta_\lambda \mathsf{L}(w_t, \lambda_t)) - w_t(\nabla_w \mathsf{L}(w_t, \lambda_t) - \delta_w \mathsf{L}(w_t, \lambda_t))\right] = 0.$$

Combining the upper bounds just derived gives:

$$\mathbb{E}\left[\max_{\lambda\in\Lambda}\mathsf{L}(w^A, \lambda) - \min_{w\in\mathcal{W}}\max_{\lambda\in\Lambda}\mathsf{L}(w, \lambda)\right] \;\leq\; \frac{2R_{\mathcal{W}}^2}{T\gamma_w} + \frac{\gamma_w(\sigma_w^2 + G_w^2)}{2} + \frac{2R_\Lambda^2}{T\gamma_\lambda} + \frac{\gamma_\lambda(\sigma_\lambda^2 + G_\lambda^2)}{2} + \frac{R_{\mathcal{W}}\sigma_w}{\sqrt{T}} + \frac{R_\Lambda\sigma_\lambda}{\sqrt{T}}.$$

Setting $\gamma_w = \frac{2R_{\mathcal{W}}}{\sqrt{T((\sigma_w^2 + G_w^2))}}$ and $\gamma_\lambda = \frac{2R_\Lambda}{\sqrt{T((\sigma_\lambda^2 + G_\lambda^2))}}$ to minimize this upper bound and using Lemma 5 completes the proof.

E.2. Proof of Lemma 2

The unbiasedness of $\delta_\lambda \mathsf{L}(w, \lambda)$ follows directly its definition. For the variance, observe that, for index $k \in [p]$, since the probability of not drawing domain k is $(1 - \frac{1}{p})$, the variance is given by the following

$$\begin{aligned}\operatorname*{Var}_k[\delta_\lambda \mathsf{L}(w, \lambda)] &= \left[1 - \frac{1}{p}\right][0 - \mathsf{L}_k(w)]^2 + \frac{1}{p}\sum_{k=1}^{p}\frac{1}{m_k}\sum_{i=1}^{m_k}[p\mathsf{L}_{k,i}(w) - \mathsf{L}_k(w)]^2 \\ &\le \left[1 - \frac{1}{p}\right]M^2 + \frac{1}{p}\sum_{k=1}^{p}\frac{1}{m_k}\sum_{i=1}^{m_k}[pM]^2 = pM^2.\end{aligned}$$

Summing over all indices from $k \in [p]$ completes the proof.

E.3. Proof of Lemma 3

The time complexity and the unbiasedness follow from the definitions. We now bound the variance. Since $\nabla_w \mathsf{L}_{k,J_k}$ is an unbiased estimate of $\nabla_w \mathsf{L}_k(w)$ and we have:

$$\operatorname{Var}[\delta_w] = \sum_{k=1}^{p}\lambda_k^2 \operatorname{Var}\left[\nabla_w \mathsf{L}_{k,J_k}(w) - \nabla_w \mathsf{L}_k(w)\right] \le \sum_{k=1}^{p}\lambda_k^2\sigma^2(w, I) \le R_\Lambda \sigma_I^2(w).$$

This completes the proof.

E.4. Proof of Lemma 4

The time complexity and the unbiasedness follow from the definitions. We now bound the variance. By definition for any w, λ,

$$\begin{aligned}\operatorname{Var}(\delta_w) &= \sum_{k=1}^{p}\frac{\lambda_k}{m_k}\sum_{j=1}^{m_k}\left(\nabla_w \mathsf{L}_{k,j}(w) - \mathsf{L}(w, \lambda)\right)^2 \\ &= \sum_{k=1}^{p}\frac{\lambda_k}{m_k}\sum_{j=1}^{m_k}\left(\nabla_w \mathsf{L}_{k,j}(w) - \mathsf{L}_k(w)\right)^2 + \sum_{k=1}^{p}\lambda_k\left(\mathsf{L}_k(w) - \mathsf{L}(w, h)\right)^2 \\ &\le \sigma_I^2(w) + \sigma_O^2(w),\end{aligned}$$

where the second equality follows from the unbiasedness of the stochastic gradients.

E.5. Comparison of PERDOMAIN and WEIGHTED stochastic gradients

For large values of p, to do a fair comparison, we need to average p independent copies of the WEIGHTED-stochastic gradient, which we refer to as p-WEIGHTED, and compare it with the PERDOMAIN-stochastic gradient. Since the variance of the average of p i.i.d. random variables is $1/p$ times the individual variance, by Lemma 4, the following holds:

$$\operatorname{Var}(p\text{-WEIGHTED}) = \frac{\sigma_I^2(w) + \sigma_O^2(w)}{p}.$$

Further, observe that $R_\Lambda = \max_{\lambda \in \Lambda}\sum_{k=1}^{p}\lambda_k^2 \ge \frac{1}{p}$. Thus,

$$\operatorname{Var}(\text{PERDOMAIN}) \ge \frac{\sigma_I^2(w)}{p}.$$

Hence, the right choice of the stochastic variance of w depends on the application. If all domains are roughly equally weighted, then we have $R(\Lambda) \approx \frac{1}{p}$ and the PERDOMAIN-variance is a more favorable choice. Otherwise, if $\sigma_O^2(w)$ is small, then the WEIGHTED-stochastic gradient is more favorable.

F. Alternative optimization algorithms for AFL

F.1. Stochastic mirror descent

In this section, we extend our STOCHASTIC-AFL algorithm to the case where a general mirror map is used, as in the mirror descent algorithm. The pseudocode of our general algorithm STOCHASTIC-MD-AFL is given in Figure 5.

```
Algorithm STOCHASTIC-MD-AFL
Initialization: $w_0 \in \mathcal{W}$ and $\lambda_0 \in \Lambda$.
Parameters: step size $\gamma_w > 0$ and $\gamma_\lambda > 0$.
For $t = 1$ to $T$:
    1. Obtain stochastic gradients: $\delta_w \mathsf{L}(w_{t-1}, \lambda_{t-1})$ and $\delta_\lambda \mathsf{L}(w_{t-1}, \lambda_{t-1})$.
    2. $w_t = \operatorname{argmin}_{w \in \mathcal{W}} \delta_w \mathsf{L}(w_{t-1}, \lambda_{t-1}) \cdot w + \frac{1}{\gamma_w} B_w(w \| w_{t-1})$.
    3. $\lambda_t = \operatorname{argmin}_{\lambda \in \Lambda} \delta_\lambda \mathsf{L}(w_{t-1}, \lambda_{t-1}) \cdot \lambda + \frac{1}{\gamma_\lambda} B_\lambda(\lambda \| \lambda_{t-1})$.
Output: $w^A = \frac{1}{T} \sum_{t=1}^T w_t$ and $\lambda^A = \frac{1}{T} \sum_{t=1}^T \lambda_t$.
```

Figure 5. Pseudocode of the STOCHASTIC-MD-AFL algorithm.

Let Φ_w be a function defined over $\operatorname{Int}(\mathcal{W})$ that is of the Legendre type (Rockafellar, 1997), that is a proper convex and differential function such that $\nabla \Phi_w$ is a one-to-one mapping from $\operatorname{Int}(\mathcal{W})$ to $\nabla \Phi_w(\operatorname{Int}(\mathcal{W}))$. Let B_w denote the Bregman divergence associated to Φ_w. For all $w, w' \in \mathcal{W}$, we have

$$B_w(w \| w') = \Phi_w(w) - \Phi_w(w') - \nabla \Phi_w(w') \cdot (w - w').$$

Similarly let Φ_λ be a Legendre-type function defined over an open set whose closure contains Λ and let B_λ denote the corresponding Bregman divergence. To simplify the notation, we use $\| \cdot \|$ to denote the norm over both w and λ, where the usage becomes clear in the context. Let $\| \cdot \|_*$ denote the corresponding dual norms. We will assume that the following properties hold.

Properties 2. *Assume that the following properties hold for the loss function* L *and sets* $\mathcal{W}$ *and* $\Lambda \subseteq \Delta_p$*:*

1. Convexity*:* $w \mapsto \mathsf{L}(w, \lambda)$ *is convex for any* $\lambda \in \Lambda$.
2. Compactness*:* $\max_{\lambda \in \Lambda} \|\lambda\| \le R_\Lambda$ *and* $\max_{w \in \mathcal{W}} \|w\| \le R_{\mathcal{W}}$*, for some* $R_\Lambda > 0$ *and* $R_{\mathcal{W}} > 0$.
3. Bounded gradients*:* $\|\nabla_w \mathsf{L}(w, \lambda)\|_* \le G_w$ *and* $\|\nabla_\lambda \mathsf{L}(w, \lambda)\|_* \le G_\lambda$ *for all* $w \in \mathcal{W}$ *and* $\lambda \in \Lambda$.
4. Stochastic variance*:* $\mathbb{E}[\|\delta_w \mathsf{L}(w, \lambda) - \nabla_w \mathsf{L}(w, \lambda)\|_*^2] \le (\sigma_w^*)^2$ *and* $\mathbb{E}[\|\delta_\lambda \mathsf{L}(w, \lambda) - \nabla_\lambda \mathsf{L}(w, \lambda)\|_*^2] \le (\sigma_\lambda^*)^2$ *for all* $w \in \mathcal{W}$ *and* $\lambda \in \Lambda$.
5. Strong convexity of Φ*: assume that* Φ_w *is* α_w*-strongly convex and* Φ_λ *is* α_λ*-strongly convex with respect to the norm* $\| \cdot \|$*. Further, assume that both* Φ_w *and* Φ_λ *are Legendre-type functions.*

With these definitions, we can now prove convergence guarantees for STOCHASTIC-MD-AFL.

Theorem 4. *[Appendix E.1] Assume that the Properties 2 hold. Then, for the step sizes* $\gamma_w = \frac{R_{\mathcal{W}} \sqrt{\alpha_w}}{\sqrt{T((\sigma_w^*)^2 + G_w^2)}}$ *and* $\gamma_\lambda = \frac{R_\Lambda \sqrt{\alpha_\lambda}}{\sqrt{T((\sigma_\lambda^*)^2 + G_\lambda^2)}}$*, the following guarantee holds for* STOCHASTIC-MD-AFL*:*

$$\mathbb{E}\left[\max_{\lambda \in \Lambda} \mathsf{L}(w^A, \lambda) - \min_{w \in \mathcal{W}} \max_{\lambda \in \Lambda} \mathsf{L}(w, \lambda)\right] \le \frac{2 R_{\mathcal{W}} \sqrt{\alpha_w((\sigma_w^*)^2 + G_w^2)}}{\sqrt{T}} + \frac{2 R_\Lambda \sqrt{\alpha_\lambda((\sigma_\lambda^*)^2 + G_\lambda^2)}}{\sqrt{T}} + \frac{R_{\mathcal{W}} \sigma_w^*}{\sqrt{T}} + \frac{R_\Lambda \sigma_\lambda^*}{\sqrt{T}}.$$

Proof. By Lemma 5, it suffices to bound $\sum_{t=1}^T \mathsf{L}(w_t, \lambda) - \mathsf{L}(w, \lambda_t)$. By (10), we can decompose this sum into three terms. The expectation of third term is zero (see proof of Theorem 4). We now bound $\sum_{t=1}^T (w_t - w) \nabla_w \mathsf{L}(w_t, \lambda_t)$. To do so, following (Mohri et al., 2018), we break step (2) of the algorithm into two equivalent steps:

$$\begin{aligned} v_t &= [\nabla \Phi_w]^{-1} (\nabla \Phi_w(w_{t-1}) - \gamma_w \delta_w \mathsf{L}(w_{t-1}, \lambda_{t-1})). \\ w_t &= \operatorname*{argmin}_{w \in \mathcal{W}} B(w \| v_t). \end{aligned}$$

We can write

$$\begin{aligned}
&\sum_{t=1}^{T}(w_t - w)\delta_w \mathsf{L}(w_t, \lambda_t) \\
&= \frac{1}{\gamma_w}\sum_{t=1}^{T}\left(\nabla\Phi_w(w_t) - \nabla\Phi_w(v_{t+1})\right)\cdot(w_t - w) && \text{(def. of } v_t\text{)} \\
&= \frac{1}{\gamma_w}\sum_{t=1}^{T}\left(B(w\|w_t) - B(w\|v_{t+1}) + B(w_t\|v_{t+1})\right) && \text{(Breg. div. Identity)} \\
&\leq \frac{1}{\gamma_w}\sum_{t=1}^{T}\left(B(w\|w_t) - B(w\|w_{t+1}) - B_w(w_{t+1}\|v_{t+1}) + B(w_t\|v_{t+1})\right) && \text{(Pythagorean ineq.)} \\
&= \frac{1}{\gamma_w}\left(B(w\|w_1) - B(w\|w_{T+1})\right) + \frac{1}{\gamma_w}\sum_{t=1}^{T}\left(-B_w(w_{t+1}\|v_{t+1}) + B(w_t\|v_{t+1})\right) && \text{(telescoping sum)} \\
&\leq \frac{B(w\|w_1)}{\gamma_w} + \frac{1}{\gamma_w}\sum_{t=1}^{T}\left(B(w_t\|v_{t+1}) - B_w(w_{t+1}\|v_{t+1})\right).
\end{aligned}$$

The second sum can be analyzed as follows:

$$\begin{aligned}
&B(w_t\|v_{t+1}) - B_w(w_{t+1}\|v_{t+1}) \\
&= \Phi_w(w_t) - \Phi_w(w_{t+1}) - \nabla\Phi_w(v_{t+1})(w_t - w_{t+1}) \\
&\leq \left(\nabla\Phi_w(w_t) - \nabla\Phi_w(v_{t+1})\right)(w_t - w_{t+1}) - \frac{\alpha_w}{2}\|w_t - w_{t+1}\|^2 && \text{(}\alpha\text{-strong convexity)} \\
&= \gamma_w \delta_w \mathsf{L}(w_t, \lambda_t)(w_t - w_{t+1}) - \frac{\alpha_w}{2}\|w_t - w_{t+1}\|^2 && \text{(def. of } v_{t+1}\text{)} \\
&\leq \gamma_w \|\delta_w \mathsf{L}(w_t, \lambda_t)\|_* \|w_t - w_{t+1}\| - \frac{\alpha_w}{2}\|w_t - w_{t+1}\|^2 && \text{(def. of dual norm)} \\
&\leq \frac{\gamma_w^2 \|\delta_w \mathsf{L}(w_t, \lambda_t)\|_*^2}{2\alpha_w} && \text{(max. of 2nd deg. eq.)} \\
&\leq \frac{\gamma_w^2\left(\|\delta_w \mathsf{L}(w_t, \lambda_t) - \nabla_w \mathsf{L}(w_t, \lambda_t)\|_*^2 + \|\nabla_w \mathsf{L}(w_t, \lambda_t)\|_*^2\right)}{\alpha_w}. && \text{(triangle ineq. and Cauchy-Schwarz)}
\end{aligned}$$

Summing the inequalities above and taking expectation yields

$$\mathbb{E}\left[\sum_{t=1}^{T}(w_t - w)\delta_w \mathsf{L}(w_t, \lambda_t)\right] \leq \frac{R_w^2}{\gamma_w} + \frac{\gamma_w((\sigma_w^*)^2 + G_w^2)}{\alpha_w}.$$

Similarly it can be shown that

$$\mathbb{E}\left[\sum_{t=1}^{T}(\lambda - \lambda_t)\delta_\lambda \mathsf{L}(w_t, \lambda_t)\right] \leq \frac{R_\lambda^2}{\gamma_\lambda} + \frac{\gamma_\lambda((\sigma_\lambda^*)^2 + G_\lambda^2)}{\alpha_w}.$$

For the second term, by the Cauchy-Schwarz inequality and Jensen's inequality, we have

$$\mathbb{E}\left[\max_{\lambda\in\Lambda}\sum_{t=1}^{T}\lambda(\nabla_\lambda \mathsf{L}(w_t, \lambda_t) - \delta_\lambda \mathsf{L}(w_t, \lambda_t))\right] \leq R_\Lambda\, \mathbb{E}\left[\|\sum_{t=1}^{T}\nabla_\lambda \mathsf{L}(w_t, \lambda_t) - \delta_\lambda \mathsf{L}(w_t, \lambda_t)\|_*\right] \leq R_\Lambda \sqrt{T}\sigma_\lambda^*.$$

Similarly, we can show that the following inequality holds:

$$\mathbb{E}\left[\max_{w\in\mathcal{W}}\sum_{t=1}^{T}w(\nabla_w \mathsf{L}(w_t, \lambda_t) - \delta_w \mathsf{L}(w_t, \lambda_t))\right] \leq R_{\mathcal{W}}\sqrt{T}\sigma_w^*.$$

Combining these inequalities, we obtain the following:

$$\begin{aligned}
&\mathbb{E}\left[\max_{\lambda\in\Lambda}\mathsf{L}(w^A, \lambda) - \min_{w\in\mathcal{W}}\max_{\lambda\in\Lambda}\mathsf{L}(w, \lambda)\right] \\
&\qquad\leq \frac{1}{T}\left[\frac{R_w^2}{\gamma_w} + \frac{\gamma_w((\sigma_w^*)^2 + G_w^2)}{\alpha_w} + \frac{R_\lambda^2}{\gamma_\lambda} + \frac{\gamma_\lambda((\sigma_\lambda^*)^2 + G_\lambda^2)}{\alpha_\lambda} + \sqrt{T}\left(R_{\mathcal{W}}\sigma_w^* + R_\Lambda\sigma_\lambda^*\right)\right].
\end{aligned}$$

Plugging in the expressions of γ_w and γ_λ completes the proof. □

```
Algorithm NON-STOCHASTIC-AFL
Initialization: $w_0$ and $\lambda_0 \in \Lambda$.
Parameters: step size $\gamma^w_{t+1} = \frac{1}{\beta_w t}$ and $\gamma^\lambda_{t+1} = \frac{1}{\beta_\lambda t}$.
For $t = 1$ to $T$:
    1. Obtain gradients: $\nabla_w \mathsf{L}(w_{t-1}, \lambda_{t-1})$ and $\nabla_\lambda \mathsf{L}(w_{t-1}, \lambda_{t-1})$.
    2. $w_t = \textsc{Project}(w_{t-1} - \gamma^w_t \nabla_w \mathsf{L}(w_{t-1}, \lambda_{t-1}), \mathcal{W})$.
    3. $\lambda_t = \textsc{Project}(\lambda_{t-1} + \gamma^\lambda_t \nabla_\lambda \mathsf{L}(w_{t-1}, \lambda_{t-1}), \Lambda)$.
Output: $w^A = \frac{1}{T}\sum_{t=1}^T w_t$ and $\lambda^A = \frac{1}{T}\sum_{t=1}^T \lambda_t$.
```

Figure 6. Pseudocode of the NON-STOCHASTIC-AFL algorithm.

F.2. Algorithm for strongly convex objectives

When the loss function is strongly convex with respect to w and strongly concave with respect to λ, conditions which often hold in the presence of regularization terms, a more favorable convergence rate of $\mathcal{O}((\log T)/T)$ can be proven for the non-stochastic algorithm NON-STOCHASTIC-AFL whose pseudocode is given in Figure 6.

Theorem 5. *Assume that the objective function is β_w-strongly convex with respect to w and β_λ-strongly concave with respect to λ, and that Properties* 1.1 *and* 1.3 *hold. Then, the following guarantee holds for* NON-STOCHASTIC AFL*:*

$$\mathbb{E}\left[\max_{\lambda\in\Lambda} \mathsf{L}(w^A, \lambda) - \min_{w\in\mathcal{W}} \max_{\lambda\in\Lambda} \mathsf{L}(w, \lambda)\right] \leq \frac{G_w^2 + G_\lambda^2}{2} \cdot \frac{1 + \log T}{T}.$$

Proof. By Lemma 5, it suffices to consider $\mathsf{L}(w_t, \lambda) - \mathsf{L}(w, \lambda_t)$. Since the function is strongly convex with respect to w and strongly concave with respect to λ,

$$\mathsf{L}(w_t, \lambda) - \mathsf{L}(w, \lambda_t) \leq (\lambda - \lambda_t)\nabla_\lambda \mathsf{L}(w_t, \lambda_t) - \beta_\lambda \|\lambda - \lambda_t\|_2^2 + (w_t - w)\nabla_w \mathsf{L}(w_t, \lambda_t) - \beta_w \|w - w_t\|_2^2.$$

We bound the term corresponding to λ,

$$\begin{aligned}
&\sum_{t=1}^T \nabla_\lambda \mathsf{L}(\lambda_t, \lambda_t) - \frac{\beta_\lambda}{2}\|\lambda - \lambda_t\|_2^2 \\
&= \sum_{t=1}^T \frac{1}{2\gamma^\lambda_{t+1}} \left(\|\lambda_t - \lambda\|_2^2 + (\gamma^\lambda_{t+1})^2 \|\nabla_\lambda \mathsf{L}(\lambda_t, \lambda_t)\|_2^2 - \|\lambda_t - \gamma^\lambda_{t+1}\nabla_\lambda \mathsf{L}(\lambda_t, \lambda_t) - \lambda\|_2^2\right) - \frac{\beta_\lambda}{2}\|\lambda - \lambda_t\|_2^2 \\
&\overset{(a)}{\leq} \sum_{t=1}^T \frac{1}{2\gamma^\lambda_{t+1}} \left(\|\lambda_t - \lambda\|_2^2 + (\gamma^\lambda_{t+1})^2 \|\nabla_\lambda \mathsf{L}(\lambda_t, \lambda_t)\|_2^2 - \|\lambda_{t+1} - \lambda\|_2^2\right) - \frac{\beta_\lambda}{2}\|\lambda - \lambda_t\|_2^2 \\
&\overset{(b)}{\leq} \frac{\beta_\lambda}{2} \sum_{t=1}^T \left((t-1)\|\lambda_t - \lambda\|_2^2 - t\|\lambda_{t+1} - \lambda\|_2^2\right) + \frac{G_\lambda^2}{2\beta_\lambda} \sum_{t=1}^T \frac{1}{t} \\
&\leq \frac{G_\lambda^2}{2\beta_\lambda}(1 + \log T),
\end{aligned}$$

where (a) follows from the property of projection and (b) follows from the definition of γ^λ_{t+1}. The following inequality can be shown in a similar way:

$$\sum_{t=1}^T (w_t - w)\nabla_w \mathsf{L}(w_t, \lambda_t) - \beta_w \|w - w_t\|_2^2 \leq \frac{G_w^2}{2\beta_w}(1 + \log T).$$

Summing up the two inequalities above and using Lemma 5 completes the proof. □

```
Algorithm Optimistic Stochastic-AFL
Initialization: $w_0$ and $\lambda_0 \in \Lambda$.
Parameters: step size $\gamma_w > 0$ and $\gamma_\lambda > 0$.
For $t = 1$ to $T$:
    1. Obtain stochastic gradients: $\delta_w \mathsf{L}(w_{t-1}, \lambda_{t-1})$ and $\delta_\lambda \mathsf{L}(w_{t-1}, \lambda_{t-1})$.
    2. $w_t = \textsc{Project}(w_{t-1} - 2\gamma_w \delta_w \mathsf{L}(w_{t-1}, \lambda_{t-1}) + \gamma_w \delta_w \mathsf{L}(w_{\max(t-2,0)}, \lambda_{\max(t-2,0)}), \mathcal{W})$.
    3. $\lambda_t = \textsc{Project}(\lambda_{t-1} + 2\gamma_\lambda \delta_\lambda \mathsf{L}(w_{t-1}, \lambda_{t-1}) - \gamma_\lambda \delta_\lambda \mathsf{L}(w_{\max(t-2,0)}, \lambda_{\max(t-2,0)}), \Lambda)$.
Output: $w_T, \lambda_T$.
```

Figure 7. Pseudocode of the Optimistic Stochastic-AFL algorithm.

F.3. Optimistic stochastic algorithm

Recently, Rakhlin & Sridharan (2013) and Daskalakis et al. (2017) gave an optimistic gradient descent algorithm for minimax optimizations. Our algorithm can also be modified to derive a stochastic optimistic algorithm, which we refer to as Optimistic-Stochastic-AFL. The pseudocode of this algorithm is also given in Figure 7. However, the convergence analysis we have presented so far does not cover this algorithm.