A. Proofs for density estimation

A.1. Proof of Lemma 1

Lemma 1. *Let the loss be a Bregman divergence B_F. Then, for any $\lambda \in \Lambda \subseteq \Delta_p$, if $h^* = \sum_{k=1}^p \lambda_k \mathcal{D}_k$ is in $\mathcal{H}$, then it is a minimizer of $h \mapsto \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big)$. If F is further strictly convex, then it is the unique minimizer.*

Proof. Fix $\lambda \in \Lambda$ such that $\sum_{k=1}^p \lambda_k \mathcal{D}_k$ is in $\mathcal{H}$. By the non-negativity of the Bregman divergence, for all h, $\mathsf{B}_F(\sum_{k=1}^p \lambda_k \mathcal{D}_k \parallel h) \geq 0$ and equality is achieved for $h = \sum_{k=1}^p \lambda_k \mathcal{D}_k$. Thus, h^* is a minimizer of $h \mapsto \mathsf{B}_F(\sum_{k=1}^p \lambda_k \mathcal{D}_k \parallel h)$. Since F is strictly convex, $h \mapsto \mathsf{B}_F(\sum_{k=1}^p \lambda_k \mathcal{D}_k \parallel h)$ is strictly convex and h^* is therefore the unique minimizer.

Now, for any hypothesis h, observe that the following difference is a constant independent of h:

$$
\begin{aligned}
&\sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big) - \mathsf{B}_F\left(\sum_{k=1}^p \lambda_k \mathcal{D}_k \parallel h\right) \qquad (12)\\
&= \sum_{k=1}^p \lambda_k \left[F(\mathcal{D}_k) - F(h) - \langle \nabla F(h), \mathcal{D}_k - h\rangle\right] - \left[F\left(\sum_{k=1}^p \lambda_k \mathcal{D}_k\right) - F(h) - \left\langle \nabla F(h), \sum_{k=1}^p \lambda_k \mathcal{D}_k - h \right\rangle\right]\\
&= \sum_{k=1}^p \lambda_k F(\mathcal{D}_k) - F\left(\sum_{k=1}^p \lambda_k \mathcal{D}_k\right).
\end{aligned}
$$

Thus, h^* is also the unique minimizer of $h \mapsto \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big)$. □

A.2. Proof of Lemma 2

Lemma 2. *Let the loss be a Bregman divergence B_F with F strictly convex and assume that $\mathrm{conv}(\{\mathcal{D}_1, ..., \mathcal{D}_p\}) \subseteq \mathcal{H}$. Observe that B_F is jointly convex in both arguments. Then, for any convex set $\Lambda \subseteq \Delta_p$, the solution of the optimization problem $\min_{h \in \mathcal{H}} \max_{\lambda \in \Lambda} \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big)$ exists and is in $\mathrm{conv}(\{\mathcal{D}_1, ..., \mathcal{D}_p\})$.*

Proof. Let $\mathcal{H}'$ is the closure of convex hull of $\mathcal{H}$. Observe that $\mathcal{H}'$ is a convex and compact set.

$$
\min_{h \in \mathcal{H}'} \max_{\lambda \in \Lambda} \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big) \leq \min_{h \in \mathcal{H}} \max_{\lambda \in \Lambda} \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big).
$$

We show that minimizer over $\mathcal{H}'$ exists and is in the $\mathrm{conv}(\{\mathcal{D}_1, \ldots, \mathcal{D}_p\})$. Since $\mathrm{conv}(\{\mathcal{D}_1, \ldots, \mathcal{D}_p\}) \subseteq \mathcal{H} \subseteq \mathcal{H}'$, the minimizer over $\mathcal{H}$ also exists and is in the $\mathrm{conv}(\{\mathcal{D}_1, \ldots, \mathcal{D}_p\})$.

Since B_F is convex with respect to its second argument, $h \mapsto \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big)$ is a convex function of h defined over the convex set $\mathcal{H}'$. Since any maximum of a convex function is also convex, $h \mapsto \max_{\lambda \in \Lambda} \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big)$ is a convex function and its minimum over the compact set $\mathcal{H}'$ exists.

We now show that the minimizer is in $\mathrm{conv}(\{\mathcal{D}_1, \ldots, \mathcal{D}_p\})$. Notice that, since $\sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big)$ is linear in λ, we have

$$
\max_{\lambda \in \Lambda} \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big) = \max_{\lambda \in \mathrm{conv}(\Lambda)} \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big).
$$

Thus, it suffices to consider the case $\Lambda \subseteq \Delta_p$. Then, since $\mathcal{H}'$ is a compact and convex set and since B_F is convex with respect to its second argument, by Sion's minimax theorem, we can write:

$$
\min_{h \in \mathcal{H}'} \max_{\lambda \in \Lambda} \sum_{k=1}^p \lambda_k \mathsf{B}_F\left(\mathcal{D}_k \parallel h\right) = \max_{\lambda \in \Lambda} \min_{h \in \mathcal{H}'} \sum_{k=1}^p \lambda_k \mathsf{B}_F\left(\mathcal{D}_k \parallel h\right).
$$

Let $\lambda^{\text{opt}} = \operatorname{argmax}_{\lambda \in \Lambda} \min_{h \in \mathcal{H}'} \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big)$ and $h^* = \sum_k \lambda_k^{\text{opt}} \mathcal{D}_k$. By assumption, $\mathrm{conv}(\{\mathcal{D}_1, \ldots, \mathcal{D}_p\})$ is included in $\mathcal{H}'$, thus h^* is in $\mathcal{H}'$ and, by Lemma 1, h^* is a minimizer of $h \mapsto \sum_{k=1}^p \lambda_k^{\text{opt}} \mathsf{B}_F(\mathcal{D}_k \parallel h)$. In view of that, if h'

is a minimizer of $h \mapsto \max_{\lambda \in \Lambda} \sum_{k=1}^p \lambda_k \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big)$ over $\mathcal{H}'$, then the following holds:

$$\begin{aligned}
\max_{\lambda} \sum_{k=1}^{p} \lambda_k \mathsf{B}_F\left(\mathcal{D}_k \parallel h'\right) &\geq \sum_{k=1}^{p} \lambda_k^{\text{opt}} \mathsf{B}_F\left(\mathcal{D}_k \parallel h'\right) && \text{(def. of } \max\text{)}\\
&\geq \sum_{k=1}^{p} \lambda_k^{\text{opt}} \mathsf{B}_F\left(\mathcal{D}_k \parallel h^*\right) && \text{(Lemma 1)}\\
&= \min_{h \in \mathcal{H}'} \sum_{k=1}^{p} \lambda_k^{\text{opt}} \mathsf{B}_F\left(\mathcal{D}_k \parallel h\right) && (h^* \text{ minimizer})\\
&= \max_{\lambda \in \Lambda} \min_{h \in \mathcal{H}'} \sum_{k=1}^{p} \lambda_k \mathsf{B}_F\left(\mathcal{D}_k \parallel h\right) && \text{(def. of } \lambda_k^{\text{opt}}\text{)}\\
&= \min_{h \in \mathcal{H}'} \max_{\lambda \in \Lambda} \sum_{k=1}^{p} \lambda_k \mathsf{B}_F\left(\mathcal{D}_k \parallel h\right). && \text{(Sion's minimax theorem)}
\end{aligned}$$

By the optimality of h', the first and last expressions in this chain of inequalities are equal, which implies the equality of all intermediate terms. In particular, this implies $\sum_{k=1}^p \lambda_k^{\text{opt}} \mathsf{B}_F(\mathcal{D}_k \parallel h') = \sum_{k=1}^p \lambda_k^{\text{opt}} \mathsf{B}_F(\mathcal{D}_k \parallel h^*)$. Since F is strictly convex, by Lemma 1, the minimizer of $h \mapsto \sum_{k=1}^p \lambda_k^{\text{opt}} \mathsf{B}_F\big(\mathcal{D}_k \parallel h\big)$ is unique and $h' = h^*$. This completes the proof. □

B. Convergence guarantee of FEDBOOST (Theorem 2)

Theorem 2. *If Properties 1 hold and* $\eta = \sqrt{\frac{\sigma}{TG^2 r_\alpha}}$, *then* α^A, *the output of* FEDBOOST *satisfies,*

$$\mathbb{E}\left[\mathsf{L}(\alpha^A) - \mathsf{L}(\alpha_{opt})\right] \leq 2\sqrt{\frac{G^2 \sigma r_\alpha}{T}} + \frac{\alpha_* M}{2T} \sum_{t=1}^{T} \sum_{k=1}^{q} \frac{\alpha_{k,t}^2}{\gamma_{k,t}}.$$

Proof. By Jensen's inequality,

$$\mathsf{L}(\alpha^A) \leq \frac{1}{T} \sum_{t=1}^{T} \mathsf{L}(\alpha_t).$$

Hence, it suffices to bound

$$\frac{1}{T} \sum_{t=1}^{T} \left(\mathsf{L}(\alpha_t) - \mathsf{L}(\alpha)\right).$$

For any t,

$$\begin{aligned}
\mathsf{L}(\alpha_t) - \mathsf{L}(\alpha) &\leq \langle \nabla \mathsf{L}(\alpha_t), \alpha_t - \alpha \rangle && \text{(convexity of } \mathsf{L}\text{)}\\
&= \langle \delta_t \mathsf{L}, \alpha_t - \alpha \rangle + \langle \nabla \mathsf{L}(\alpha_t) - \delta_t \mathsf{L}, \alpha_t - \alpha \rangle\\
&= \frac{1}{\eta} \langle \nabla F(\alpha_t) - \nabla F(v_{t+1}), \alpha_t - \alpha \rangle + \langle \nabla \mathsf{L}(\alpha_t) - \delta_t \mathsf{L}, \alpha_t - \alpha \rangle && \text{(def. of } v_{t+1}\text{)}\\
&= \frac{1}{\eta} \left(\mathsf{B}_F(\alpha \parallel \alpha_t) + \mathsf{B}_F(\alpha_t \parallel v_{t+1}) - \mathsf{B}_F(\alpha \parallel v_{t+1})\right) + \langle \nabla \mathsf{L}(\alpha_t) - \delta_t \mathsf{L}, \alpha_t - \alpha \rangle && \text{(Bregman div. def.)}\\
&\leq \frac{1}{\eta} \left(\mathsf{B}_F(\alpha \parallel \alpha_t) + \mathsf{B}_F(\alpha_t \parallel v_{t+1}) - \mathsf{B}_F(\alpha \parallel \alpha_{t+1}) - \mathsf{B}_F(\alpha_{t+1} \parallel v_{t+1})\right) && \text{(13a)}\\
&+ \langle \nabla \mathsf{L}(\alpha_t) - \delta_t \mathsf{L}, \alpha_t - \alpha \rangle, && \text{(13b)}
\end{aligned}$$

where the last inequality follows because $\mathsf{B}_F(\alpha \parallel v_{t+1}) \geq \mathsf{B}_F(\alpha \parallel \alpha_{t+1}) + \mathsf{B}_F(\alpha_{t+1} \parallel v_{t+1})$ by the generalized

Pythagorean inequality. For the first term (13a), summing over t gives the following telescoping sum,

$$\begin{aligned}
&\sum_{t=1}^{T}(\mathsf{B}_F(\alpha \parallel \alpha_t) + \mathsf{B}_F(\alpha_t \parallel v_{t+1}) - \mathsf{B}_F(\alpha \parallel \alpha_{t+1}) - \mathsf{B}_F(\alpha_{t+1} \parallel v_{t+1})) \qquad (14)\\
&\qquad = \mathsf{B}_F(\alpha \parallel \alpha_1) - \mathsf{B}_F(\alpha \parallel \alpha_{T+1}) + \sum_{t=1}^{T} \mathsf{B}_F(\alpha_t \parallel v_{t+1}) - \mathsf{B}_F(\alpha_{t+1} \parallel v_{t+1})\\
&\qquad \le \mathsf{B}_F(\alpha \parallel \alpha_1) + \sum_{t=1}^{T} (\mathsf{B}_F(\alpha_t \parallel v_{t+1}) - \mathsf{B}_F(\alpha_{t+1} \parallel v_{t+1}))\,.
\end{aligned}$$

Now consider the summation term:

$$\begin{aligned}
&\mathsf{B}_F(\alpha_t \parallel v_{t+1}) - \mathsf{B}_F(\alpha_{t+1} \parallel v_{t+1}) = F(\alpha_t) - F(\alpha_{t+1}) - \langle \nabla F(v_{t+1}), \alpha_t - \alpha_{t+1} \rangle\\
&\qquad \le \langle \nabla F(\alpha_t), \alpha_t - \alpha_{t+1} \rangle - \frac{\sigma}{2}\|\alpha_t - \alpha_{t+1}\|^2 - \langle \nabla F(v_{t+1}), \alpha_t - \alpha_{t+1} \rangle \qquad \text{(strong convexity of } F\text{)}\\
&\qquad = \langle \nabla F(\alpha_t) - \nabla F(v_{t+1}), \alpha_t - \alpha_{t+1} \rangle - \frac{\sigma}{2}\|\alpha_t - \alpha_{t+1}\|^2\\
&\qquad = \eta \langle \delta_t \mathsf{L}, \alpha_t - \alpha_{t+1} \rangle - \frac{\sigma}{2}\|\alpha_t - \alpha_{t+1}\|^2 \qquad \text{(def. of } v_{t+1}\text{)}\\
&\qquad \le \eta \|\delta_t \mathsf{L}\|_* \|\alpha_t - \alpha_{t+1}\| - \frac{\sigma}{2}\|\alpha_t - \alpha_{t+1}\|^2 \qquad \text{(Cauchy-Schwarz ineq.)}\\
&\qquad \le \frac{\eta^2 \|\delta_t \mathsf{L}\|_*^2}{2\sigma}. \qquad (15)
\end{aligned}$$

Combining the above inequalities,

$$\begin{aligned}
\sum_{t=1}^{T} (\mathsf{L}(\alpha_t) - \mathsf{L}(\alpha)) &\le \frac{1}{\eta}\mathsf{B}_F(\alpha \parallel \alpha_1) + \sum_{t=1}^{T} \left(\frac{\eta \|\delta_t \mathsf{L}\|_*^2}{2\sigma} + \langle \nabla \mathsf{L}(\alpha_t) - \delta_t \mathsf{L}, \alpha_t - \alpha \rangle \right)\\
&\le \frac{1}{\eta}\mathsf{B}_F(\alpha \parallel \alpha_1) + \frac{\eta G^2 T}{2\sigma} + \sum_{t=1}^{T} (\langle \nabla \mathsf{L}(\alpha_t) - \delta_t \mathsf{L}, \alpha_t - \alpha \rangle)\,.
\end{aligned}$$

We now bound (13b) in expectation, the inner product term in the above equation. Denote by $\nabla_t \mathsf{L}(\cdot) := \sum_{j \in S_t} \frac{m_j}{m} \nabla \mathsf{L}_j(\cdot)$, where $m = \sum_{j \in S_t} m_j$. Taking the expectation over $j \in S_t$,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \langle \nabla \mathsf{L}(\alpha_t) - \delta_t \mathsf{L}, \alpha_t - \alpha \rangle \right] &= \sum_{t=1}^{T} \langle \nabla \mathsf{L}(\alpha_t) - \mathbb{E}\left[\delta_t \mathsf{L}\right], \alpha_t - \alpha \rangle \qquad (16)\\
&= \sum_{t=1}^{T} \langle \nabla \mathsf{L}(\alpha_t) - \mathbb{E}\left[\nabla_t \mathsf{L}(\tilde{\alpha}_t)\right], \alpha_t - \alpha \rangle\\
&\le \sum_{t=1}^{T} \|\nabla \mathsf{L}(\alpha_t) - \mathbb{E}\left[\nabla_t \mathsf{L}(\tilde{\alpha}_t)\right]\|_* \|\alpha_t - \alpha\| \qquad \text{(Cauchy-Schwarz ineq.)}\\
&\le \sum_{t=1}^{T} \|\nabla \mathsf{L}(\alpha_t) - \mathbb{E}\left[\nabla_t \mathsf{L}(\tilde{\alpha}_t)\right]\|_* \alpha_*. \qquad \text{(by Prop. 1.2.)}
\end{aligned}$$

$$(17)$$

To understand $\mathbb{E}[\nabla_t \mathsf{L}(\tilde{\alpha}_t)]$, we use Taylor's Theorem in several variables (Folland, 2010). Let $f = \nabla_t \mathsf{L}$. Expanding $f(\tilde{\alpha}_t)$ about α_t,

$$f(\tilde{\alpha}_t) - f(\alpha_t) = \nabla f(\alpha_t)(\tilde{\alpha}_t - \alpha_t) + R_1(\tilde{\alpha}_t - \alpha_t),$$

where $R_1(\cdot)$ is the reminder term that can be bounded as

$$\|R_1(\tilde{\alpha}_t - \alpha_t)\|_* \leq \frac{M}{2!}\|\tilde{\alpha}_t - \alpha_t\|_2^2,$$

with $\|\nabla^2 f(\cdot)\| \leq M$. Taking expectation over $\tilde{\alpha}_t$ and using the fact that $\mathbb{E}[\tilde{\alpha}_t] = \alpha_t$, we get

$$\begin{aligned}
\| \mathbb{E}[f(\tilde{\alpha}_t)] - f(\alpha_t)\| &\leq \frac{M}{2} \mathbb{E} \|\tilde{\alpha}_t - \alpha_t\|_2^2 \\
&= \frac{M}{2} \sum_{k=1}^{q} \mathbb{E} \left\| \frac{\alpha_{k,t} \mathbf{1}_{k,t}}{\gamma_{k,t}} - \alpha_{k,t} \right\|_2^2 \\
&= \frac{M}{2} \sum_{k=1}^{q} \alpha_{k,t}^2 \, \mathbb{E} \left\| \frac{\mathbf{1}_{k,t}}{\gamma_{k,t}} - 1 \right\|_2^2 \\
&= \frac{M}{2} \sum_{k=1}^{q} \alpha_{k,t}^2 \left(\frac{1 - \gamma_{k,t}}{\gamma_{k,t}} \right) \\
&\leq \frac{M}{2} \sum_{k=1}^{q} \frac{\alpha_{k,t}^2}{\gamma_{k,t}}.
\end{aligned}$$

Combining the resulting inequalities gives

$$\mathsf{L}(\alpha^A) - \mathsf{L}(\alpha) \leq \frac{1}{\eta T} \mathsf{B}_F(\alpha \parallel \alpha_1) + \frac{\eta G^2}{2\sigma} + \frac{\alpha_* M}{2T} \sum_{t=1}^{T} \sum_{k=1}^{q} \frac{\alpha_{k,t}^2}{\gamma_{k,t}}.$$

Choosing the learning rate yields the theorem. $\square$

C. Convergence guarantee for AFLBoost (Theorem 3)

Theorem 3. *Let Properties 1 and 2 hold. Let* $\eta_\lambda = \sqrt{\frac{\sigma}{T G_\lambda^2 r_\lambda}}$ *and* $\eta_\alpha = \sqrt{\frac{\sigma}{T G_\alpha^2 r_\alpha}}$. *Let* α^A *be the output of* AFLBoost. *If* $\gamma_{k,t}$ *is given by 8, then* $\mathbb{E}[\max_{\lambda \in \Lambda} \mathsf{L}(\alpha^A, \lambda) - \min_{\alpha \in \Delta_q} \max_{\lambda \in \Lambda} \mathsf{L}(\alpha, \lambda)]$ *is at most*

$$4\sqrt{\frac{G_\alpha^2(\sigma r_\alpha + \alpha_*)}{T}} + 4\sqrt{\frac{G_\lambda^2(\sigma r_\lambda + \lambda_*)}{T}} + \frac{M(\lambda_* + \alpha_*)}{C}.$$

Proof. By Mohri et al. (2019)[Lemma 5], it suffices to bound

$$\frac{1}{T} \max_{\substack{\lambda \in \Lambda; \\ \alpha \in \Delta_q}} \{ \sum_{t=1}^{T} \mathsf{L}(\alpha_t, \lambda) - \mathsf{L}(\alpha, \lambda_t) \}. \tag{18}$$

Consider the following inequalities:

$$\begin{aligned}
\mathsf{L}(\alpha_t, \lambda) - \mathsf{L}(\alpha, \lambda_t) &= \mathsf{L}(\alpha_t, \lambda) - \mathsf{L}(\alpha_t, \lambda_t) + \mathsf{L}(\alpha_t, \lambda_t) - \mathsf{L}(\alpha, \lambda_t) \\
&\leq \langle \nabla_\lambda \mathsf{L}(\alpha_t, \lambda_t), \lambda - \lambda_t \rangle + \langle \nabla_\alpha \mathsf{L}(\alpha_t, \lambda_t), \alpha_t - \alpha \rangle && \text{(convexity of } \mathsf{L}) \\
&= \langle \delta_{\lambda,t} \mathsf{L}, \lambda - \lambda_t \rangle + \langle \delta_{\alpha,t} \mathsf{L}, \alpha_t - \alpha \rangle \\
&+ \langle \nabla_\lambda \mathsf{L}(\alpha_t, \lambda_t) - \delta_{\lambda,t} \mathsf{L}, \lambda - \lambda_t \rangle + \langle \nabla_\alpha \mathsf{L}(\alpha_t, \lambda_t) - \delta_{\alpha,t} \mathsf{L}, \alpha_t - \alpha \rangle
\end{aligned}$$

Given these inequalities, we can bound (18) using the sub-additive property of $\max$ on the previous inequality as follows:

$$\max_{\substack{\lambda\in\Lambda;\\ \alpha\in\Delta_q}}\{\sum_{t=1}^{T}\mathsf{L}(\alpha_t,\lambda)-\mathsf{L}(\alpha,\lambda_t)\}$$

$$\leq \max_{\substack{\lambda\in\Lambda;\\ \alpha\in\Delta_q}}\sum_{t=1}^{T}\{\langle\delta_{\lambda,t}\mathsf{L},\lambda-\lambda_t\rangle+\langle\delta_{\alpha,t}\mathsf{L},\alpha_t-\alpha\rangle\} \tag{19a}$$

$$+\max_{\substack{\lambda\in\Lambda;\\ \alpha\in\Delta_q}}\sum_{t=1}^{T}\{\langle\lambda,\nabla_\lambda\mathsf{L}(\alpha_t,\lambda_t)-\delta_{\lambda,t}\mathsf{L}\rangle+\langle\alpha,\nabla_\alpha\mathsf{L}(\alpha_t,\lambda_t)-\delta_{\alpha,t}\mathsf{L}\rangle\} \tag{19b}$$

$$+\sum_{t=1}^{T}\langle\lambda_t,\nabla_\lambda\mathsf{L}(\alpha_t,\lambda_t)-\delta_{\lambda,t}\mathsf{L}\rangle+\langle\alpha_t,\nabla_\alpha\mathsf{L}(\alpha_t,\lambda_t)-\delta_{\alpha,t}\mathsf{L}\rangle, \tag{19c}$$

which we will bound in three parts. Consider the first sub-equation (19a): similarly in arriving at (13a), it follows by definition of w_{t+1},v_{t+1} that

$$\begin{aligned}\langle\delta_{\lambda,t}\mathsf{L},\lambda-\lambda_t\rangle+\langle\delta_{\alpha,t}\mathsf{L},\alpha_t-\alpha\rangle\leq{}&\frac{1}{\eta_\lambda}(\mathsf{B}_F(\lambda\parallel\lambda_t)+\mathsf{B}_F(\lambda_t\parallel w_{t+1})-\mathsf{B}_F(\lambda\parallel\lambda_{t+1})-\mathsf{B}_F(\lambda_{t+1}\parallel w_{t+1}))\\&+\frac{1}{\eta_\alpha}(\mathsf{B}_F(\alpha\parallel\alpha_t)+\mathsf{B}_F(\alpha_t\parallel v_{t+1})-\mathsf{B}_F(\alpha\parallel\alpha_{t+1})-\mathsf{B}_F(\alpha_{t+1}\parallel v_{t+1})).\end{aligned}$$

Summing over t, this gives the following by similar argument as in (14) for all λ,α:

$$\begin{aligned}\sum_{t=1}^{T}\langle\delta_{\lambda,t}\mathsf{L},\lambda-\lambda_t\rangle+\langle\delta_{\alpha,t}\mathsf{L},\alpha_t-\alpha\rangle\leq{}&\frac{1}{\eta_\lambda}(\mathsf{B}_F(\lambda\parallel\lambda_1)+\sum_{t=1}^{T}\mathsf{B}_F(\lambda_t\parallel w_{t+1})-\mathsf{B}_F(\lambda_{t+1}\parallel w_{t+1}))\\&+\frac{1}{\eta_\alpha}(\mathsf{B}_F(\alpha\parallel\alpha_1)+\sum_{t=1}^{T}\mathsf{B}_F(\alpha_t\parallel v_{t+1})-\mathsf{B}_F(\alpha_{t+1}\parallel v_{t+1}))\end{aligned}$$

In view of the inequality resulting from (15), for all λ,α, this is bounded by

$$\begin{aligned}\sum_{t=1}^{T}\langle\delta_{\lambda,t}\mathsf{L},\lambda-\lambda_t\rangle+\langle\delta_{\alpha,t}\mathsf{L},\alpha_t-\alpha\rangle&\leq\frac{1}{\eta_\lambda}\mathsf{B}_F(\lambda\parallel\lambda_1)+\frac{1}{\eta_\alpha}\mathsf{B}_F(\alpha\parallel\alpha_1)+\sum_{t=1}^{T}\frac{\eta_\lambda^2\|\delta_{\lambda,t}\mathsf{L}\|_*^2+\eta_\alpha^2\|\delta_{\alpha,t}\mathsf{L}\|_*^2}{2\sigma}\\&=\frac{1}{\eta_\lambda}\mathsf{B}_F(\lambda\parallel\lambda_1)+\frac{1}{\eta_\alpha}\mathsf{B}_F(\alpha\parallel\alpha_1)+\frac{T\left(\eta_\lambda G_\lambda^2+\eta_\alpha G_\alpha^2\right)}{2\sigma}.\end{aligned}$$

Next, we proceed with the bound for third sub-equation (19c) in expectation via similar argument followed to arrive at (15):

$$\begin{aligned}&\mathbb{E}[\sum_{t=1}^{T}\langle\lambda_t,\nabla_\lambda\mathsf{L}(\alpha_t,\lambda_t)-\delta_{\lambda,t}\mathsf{L}\rangle+\langle\alpha_t,\nabla_\alpha\mathsf{L}(\alpha_t,\lambda_t)-\delta_{\alpha,t}\mathsf{L}\rangle]\\&\qquad=\sum_{t=1}^{T}\langle\lambda_t,\nabla_\lambda\mathsf{L}(\alpha_t,\lambda_t)-\mathbb{E}[\nabla_{t,\lambda}\mathsf{L}(\tilde{\alpha}_t,\lambda_t)]\rangle+\langle\alpha_t,\nabla_\alpha\mathsf{L}(\tilde{\alpha}_t,\lambda_t)-\mathbb{E}[\nabla_{t,\alpha}\mathsf{L}(\tilde{\alpha}_t,\lambda_t)]\rangle,\end{aligned}$$

where $\nabla_{t,\lambda}\mathsf{L}(\cdot):=\sum_{j\in S_t}\frac{m_j}{m}\nabla_\lambda\mathsf{L}_j(\cdot)$, and similarly for $\nabla_{t,\alpha}\mathsf{L}(\cdot)$. Similar to the proof of (15) and (9), it can be shown that

$$\sum_{t=1}^{T}\langle\lambda_t,\nabla_\lambda\mathsf{L}(\alpha_t,\lambda_t)-\mathbb{E}[\nabla_{t,\lambda}]\rangle\leq\frac{MT\lambda_*}{2C}.$$

Similarly,

$$\mathbb{E}[\sum_{t=1}^{T}\langle\alpha_t,\nabla_\alpha\mathsf{L}(\alpha_t,\lambda_t)-\delta_{\alpha,t}\mathsf{L}\rangle]\leq\frac{MT\alpha_*}{2C}.$$

Combining the two bounds, we have

$$\mathbb{E}[\sum_{t=1}^{T}\langle \lambda_t, \nabla_\lambda \mathsf{L}(\alpha_t, \lambda_t) - \delta_{\lambda,t}\mathsf{L}\rangle + \langle \alpha_t, \nabla_\alpha \mathsf{L}(\alpha_t, \lambda_t) - \delta_{\alpha,t}\mathsf{L}\rangle] \le \frac{MT(\alpha_* + \lambda_*)}{2C}.$$

We now consider the second sub-equation term (19b), focusing on the first summand with the max over λ and bound this by the Cauchy-Schwarz inequality, then Jensen's inequality:

$$\begin{aligned}
&\mathbb{E}[\max_{\lambda\in\Lambda}\{\sum_{t=1}^{T}\langle \lambda, \nabla_\lambda \mathsf{L}(\alpha_t, \lambda_t) - \delta_{\lambda,t}\mathsf{L}\rangle\}]\\
&\le \mathbb{E}[\max_{\lambda\in\Lambda}\{\sum_{t=1}^{T}\langle \lambda, \nabla_\lambda \mathsf{L}(\alpha_t, \lambda_t) - \mathbb{E}[\delta_{\lambda,t}\mathsf{L}]\rangle\}] + \mathbb{E}[\max_{\lambda\in\Lambda}\{\sum_{t=1}^{T}\langle \lambda, \delta_{\lambda,t}\mathsf{L} - \mathbb{E}[\delta_{\lambda,t}\mathsf{L}]\rangle\}]\\
&\le \frac{MT\lambda_*}{2C} + \lambda_* G_\lambda \sqrt{T},
\end{aligned}$$

where λ_* denotes the $\max$ over the compact set Λ. Similarly, we can obtain the following inequality:

$$\mathbb{E}[\max_{\alpha\in\Delta_q}\sum_{t=1}^{T}\langle \alpha, \nabla_\alpha \mathsf{L}(\alpha_t, \lambda_t) - \delta_{\alpha,t}\mathsf{L}\rangle] \le \frac{MT\alpha_*}{2C} + \alpha_* G_\alpha \sqrt{T}.$$

Thus, combining the inequalities gives

$$\max_{\substack{\lambda\in\Lambda;\\ \alpha\in\Delta_q}}\sum_{t=1}^{T}\{\langle \lambda, \nabla_\lambda \mathsf{L}(\alpha_t, \lambda_t) - \delta_{\lambda,t}\mathsf{L}\rangle + \langle \alpha, \nabla_\alpha \mathsf{L}(\alpha_t, \lambda_t) - \delta_{\alpha,t}\mathsf{L}\rangle\} = \frac{MT(\lambda_* + \alpha_*)}{2C} + \alpha_* G_\alpha\sqrt{T} + \lambda_* G_\lambda \sqrt{T}.$$

Combining the bounds for (19a), (19b), and (19c), the following bound holds:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\eta_\alpha \mathsf{L}(\alpha_t, \lambda) - \eta_\lambda \mathsf{L}(\alpha, \lambda_t)\\
&\qquad\le \frac{1}{T}\left(\frac{\mathsf{B}_F(\lambda \,\|\, \lambda_1)}{\eta_\lambda} + \frac{\mathsf{B}_F(\alpha \,\|\, \alpha_1)}{\eta_\alpha}\right) + \frac{T\left(\eta_\lambda G_\lambda^2 + \eta_\alpha G_\alpha^2\right)}{2\sigma} + \frac{M(\lambda_* + \alpha_*)}{C} + \frac{\alpha_* G_\alpha + \lambda_* G_\lambda}{\sqrt{T}}.
\end{aligned}$$

$\square$

D. Additional density estimation experiments on synthetic data

Continuing the experimental validation of FEDBOOST as described in 5.1, we examine the effect of modulating the communication budget C on a density estimation task using the same setup as before, but with a power-law distributed synthetic dataset with parameter $p = 1000$.

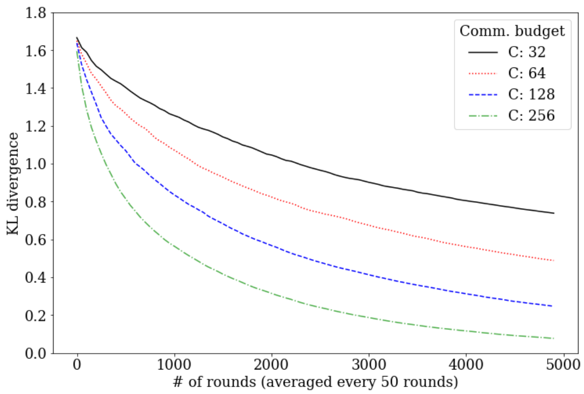

Figure 6. Comparison of loss curves as a function of C using *uniform sampling* in density estimation on synthetic data.

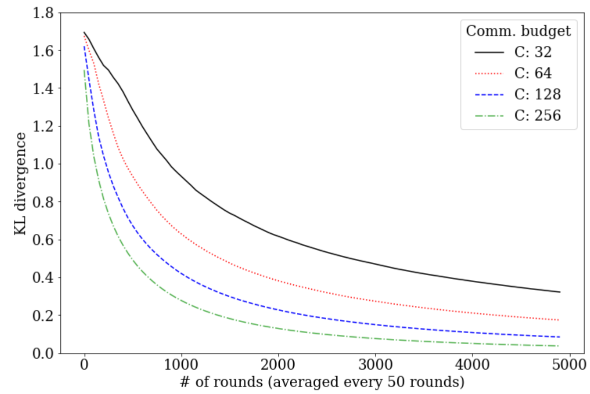

Figure 7. Comparison of convergence as C varies using *weighted random sampling* for density estimation.

We use a hand-tuned step size $\eta = 0.001$ for all values of C, and include ℓ_1 regularization in the experiment using *weighted random sampling* (Fig. 7). The experimental setup is otherwise the same for both Fig. 6 and 7. Across all values of C, *weighted random sampling* of the h_k achieves lower loss than using *uniform sampling*, which validates that using weighted random sampling reduces the communication-dependent term of FEDBOOST.