Optimal Model Averaging:
Towards Personalized Collaborative Learning

Felix Grimberg [1]   **Mary-Anne Hartley** [1]   **Sai P. Karimireddy** [1]   **Martin Jaggi** [1]

Abstract

In federated learning, differences in the data or objectives between the participating nodes motivate approaches to train a personalized machine learning model for each node. One such approach is weighted averaging between a locally trained model and the global model. In this theoretical work, we study weighted model averaging for arbitrary scalar mean estimation problems under minimal assumptions on the distributions. In a variant of the bias-variance trade-off, we find that there is always some positive amount of model averaging that reduces the expected squared error compared to the local model, provided only that the local model has a non-zero variance. Further, we quantify the (possibly negative) benefit of weighted model averaging as a function of the weight used and the optimal weight. Taken together, this work formalizes an approach to quantify the value of personalization in collaborative learning and provides a framework for future research to test the findings in multivariate parameter estimation and under a range of assumptions.

1. Introduction

Collaborative learning refers to the setting where several agents, each having access to their own personal training data, collaborate in hopes of learning a better model. In most real applications, the data distributions are different for each agent. Each agent therefore faces a key decision, in order to obtain the highest quality model: Should they ignore their peers and simply train a model on their local data alone—or should they collaborate to potentially benefit from additional training data from other agents?

This fundamental question is at the center of two areas that are currently of high interest in both research and industry applications, namely learning personalized models as well as federated and decentralized learning.

Federated Learning. In federated learning (FL) (Konečný et al., 2016), training data is distributed over several agents or locations. For instance, these could be several hospitals collaborating on a clinical trial, or billions of mobile phones involved in training a next-word prediction model for a virtual keyboard application. The purpose of FL is to train a global model on the union of all agents' individual data. The training is coordinated by a central server, while the agents' local data never leaves its device of origin. Owing to the data locality, FL has become the most prominent collaborative learning approach in recent years towards privacy-preserving machine learning (Kairouz et al., 2019; Li et al., 2020). *Decentralized learning* refers to the analogous setting without a central server, where agents communicate peer-to-peer during training, see e.g. (Nedic, 2020).

Personalization. When each agent has a different data distribution[1] (and thus a different learning task), a "one model fits all" approach leads to poor accuracy on individual agents. Instead, a given global model (such as from FL) needs to be personalized, e.g., by additional training on the local data of our agent. Prominent approaches to address this important problem of statistical heterogeneity are additional local training or fine-tuning (Wang et al., 2019), or weighted averaging between a global model and a locally trained model, during or after training (Mansour et al., 2020; Deng et al., 2020).

The Question. In this paper, we address weighted averaging of several models, asking: *When can a weighted model average outperform an individual agent's local model? How much can we gain by using it? How much can we lose?* We aim to answer this while making only minimal assumptions on the data distributions.

In the context of FL, an answer to these questions serves the users to determine under what conditions FL should

---

[1]School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland. Correspondence to: Felix Grimberg <felix.grimberg@hotmail.com>.

---

[1]   See Kairouz et al. (2019, Section 3.1) for common types of violations of independence and identical distribution in FL.

be preferred to independently training a local model. It is also of interest to the FL server to identify and potentially reward participants with high contributions to the model quality. In decentralized learning, it can be used by agents to select most compatible peers during training (Grimberg et al., 2020). The question also naturally extends to model personalization, as in Deng et al. (2020) who ask *"when is personalization better?"* and *"what degree of personalization is best?"* in the context of model interpolation.

Contributions. In this work, we analyse the linear combination of two models for arbitrary scalar mean estimation problems. Given a local empirical mean $\bar{X}$ and some other empirical mean $\bar{Y}$, we ask whether the *weighted model average* $(1 - \alpha)\bar{X} + \alpha\bar{Y}$ is a better estimator of the *local true mean* $\mathbb{E}[\bar{X}]$, than $\bar{X}$ itself. For instance, $\bar{Y}$ could be a global model obtained through federated learning.

We calculate the error of the weighted model average with respect to the local true mean, taking the expectation over $\bar{X}$ and $\bar{Y}$. We find the optimal weight $\alpha^\star$ to minimize this error, showing that the error of the optimally weighted average is reduced by a fraction $\alpha^\star$ (equal to the weight itself), compared to the error of $\bar{X}$.

In a variant of the bias-variance trade-off, we find that there is always some positive amount $\alpha^\star > 0$ of model averaging that reduces the error compared to the local model $\bar{X}$ (provided that $\bar{X}$ has a non-zero variance). Recognizing that the optimal weight depends on quantities that are likely unknown to the experimenter, we quantify the error of a sub-optimally weighted model average with weight α. We show that the error is better than the local model's if $\alpha < 2\alpha^\star$, and that even a small weight $\alpha < \alpha^\star$ can lead to a relatively large improvement. On the other hand, the error is worse than the local model's if $\alpha > 2\alpha^\star$. Thus, choosing an exceedingly large weight can be harmful in situations where $2\alpha^\star < 1$. This could easily happen in practice, if α is chosen based on the observed data in the presence of a large sampling bias.

We introduce our model in Section 3, showing how it relates to practical use cases and reflecting on the assumptions underlying our results. We prove our main results in Section 4 and visualize and discuss them in Section 5. Here, we also interpret weighted model averaging as a general form of shrinkage and explain how our results compare to recent related work. We conclude with a brief summary of our results and pointers to open problems in Section 6.

2. Related Work

Model Personalization. A properly identified source of heterogeneity between the agents can sometimes be addressed by incorporating context information as additional features in the ML model: location, time, user demographic,

etc. Several other solutions have been proposed to tackle unidentified sources of heterogeneity:

Local fine-tuning, wherein the fully trained global model is adapted to the local context by taking a number of local stochastic gradient descent (SGD) steps (Wang et al., 2019).

Multi-task learning and clustering, to group agents and train one model per cluster. While agents can often be clustered by their context or geolocalisation, Mansour et al. (2020) propose an expectation-maximization-type algorithm to jointly optimize the models and the clustering.

Gradient-based personalization methods, that learn a personalized model by using variants of gradient similarity measures to adapt the weight of local and global models (Grimberg et al., 2020; Fallah et al., 2020).

Joint local and global learning, where a global model and a local model are trained simultaneously (Deng et al., 2020; Mansour et al., 2020; Fallah et al., 2020), for example using model-agnostic meta-learning (Kairouz et al., 2019).

Merging already trained models is also viable using approaches such as model fusion (Singh & Jaggi, 2020) or distillation (Lin et al., 2020), however these again come without theoretical guarantees on the quality of the resulting merged model.

Theoretical Analysis of Weighted Model Averaging. Donahue & Kleinberg (2020) investigate FL from a game-theory point of view, to test whether self-interested players (i.e., agents) have an incentive to join an FL task. They analyze the formation of clusters *(stable coalitions)* of players in a linear regression task with specific assumptions on the data generation process. Once allowing each player to use a weighted average of their local model with their cluster's global model, the *grand coalition* (i.e. a federation of all agents) is weakly preferred by all players over their local model. Further, if players can weight each other player's local model individually rather than selecting a single weight for their cluster's model, the *grand coalition* becomes core stable, where no other coalition C is preferred over the *grand coalition* by all players in C.

Thus, weighted model averaging seems more promising than clustering-based approaches in the setting under consideration. We expand on this analysis of weighted model averaging, proving that the results about the optimal model averaging weight hold even under minimal assumptions on the data generation process. The work of Donahue & Kleinberg (2020) is based on unrealistic assumptions of identical variance and similar means across players. In contrast, we allow arbitrary finite means and variances, only requiring a positive variance for the local model. Rather, our analysis is limited to a one-dimensional mean estimation problem with two players ($D = 1$, $M = 2$). As shown in Section 5.5, our

results are equivalent to theirs when our respective assumptions are applied jointly.

The Bias-Variance Trade-Off. Weighted model averaging is a form of bias-variance trade-off: It aims to leverage the vast quantity of data on the network to reduce the model's generalization error, at the expense of increased bias if the local and global distributions match poorly. Theoretical research on the bias-variance trade-off dates back to the surprising results of James & Stein (1961), who constructed a biased empirical estimator that provably dominates the maximum likelihood estimator for a specific higher-dimensional mean estimation problem (cf. Section 5.4). Today, we would call this a *shrinkage* estimator.

Despite their long history, shrinkage estimators are still relevant and can be found in recent applications. For instance, Su et al. (2020) recently investigated how to shrink importance weights to reduce the mean squared error of doubly-robust estimators in off-policy evaluation. In Section 5.4, we show that our analysis of weighted model averaging includes shrinkage as a special case, when $\bar{Y}$ is set to a constant value, such as 0.

3. Model and Assumptions

3.1. Setup

To introduce the theoretical setting, we suppose that an agent a_X has drawn n_X independent samples from the random variable X, and that a_X wishes to estimate the unknown true mean $\mu_X \triangleq \mathbb{E}[X]$. For instance, a_X could be a clinical researcher interested in accurately estimating the effect size of a new treatment from the n_X patients participating in a trial at their hospital.

Agent a_X can construct an empirical estimator for μ_X based on its own n_X samples. For instance, a_X could use the local empirical mean $\bar{X}$ from Definition 1, which is unbiased and consistent. Alternatively, agent a_X can enlist the help of another agent, a_Y, who has calculated the empirical mean $\bar{Y}$ of n_Y samples drawn from a different random variable Y. This allows a_X to estimate μ_X by a weighted average of $\bar{X}$ and $\bar{Y}$, denoted as $\bar{\mu}_\alpha$, where $\alpha \in [0,1]$ is the weight of $\bar{Y}$ (Definition 2).

Definition 1 (Empirical mean). *Let $\{x_1, \ldots, x_{n_X}\} \overset{i.i.d.}{\sim} X$, $\{y_1, \ldots, y_{n_Y}\} \overset{i.i.d.}{\sim} Y$. We recall the empirical means $\bar{X}$ and $\bar{Y}$:*

$$\bar{X} \triangleq \frac{1}{n_X} \sum_{i=1}^{n_X} x_i, \quad \bar{Y} \triangleq \frac{1}{n_Y} \sum_{i=1}^{n_Y} y_i$$

Definition 2 (Weighted average). *We define the weighted model average $\bar{\mu}_\alpha$ of the two means as:*

$$\bar{\mu}_\alpha \triangleq (1 - \alpha) \bar{X} + \alpha \bar{Y}$$

Examples of Helper Agent a_Y. In our example of a clinical trial, the helper agent a_Y could be another researcher conducting a trial of a similar treatment in a different location. Alternatively, $\bar{Y}$ could also be a global model obtained through federated learning by several other hospitals conducting treatment trials. As we do not make any assumptions on Y, it could also simply be an arbitrary constant (cf. Section 5.4). For instance, picking $Y = 0$ results in the shrinkage estimator $\bar{\mu}_\alpha = (1 - \alpha) \bar{X}$.

3.2. Problem Definition

We assume that a_X wants to use the estimator whose *expected squared error* (ESE) with respect to μ_X is minimal. *Optimality* is therefore understood throughout this paper in terms of the ESE. Note that the estimator $\bar{\mu}_\alpha$ is biased for $\alpha > 0$ unless $\mathbb{E}[Y] = \mu_X$, see Definition 3.

Definition 3 (Expected squared error). *Recall that $\mu_X \triangleq \mathbb{E}[X]$. By e_α, we denote the ESE of $\bar{\mu}_\alpha$ w.r.t. μ_X.*

$$e_\alpha = \mathbb{E}\left[(\bar{\mu}_\alpha - \mu_X)^2\right]$$

This focus on the estimator's ESE implies that a_X is equally concerned about underestimating the mean, as about overestimating it—which need not be the case in practice. Furthermore, by focusing on the squared error, we assume that a_X prefers a high probability of incurring a relatively small error, than an n times lower probability of incurring an n times larger error.

The task of a_X reduces to selecting an optimal weight $\alpha^\star$ (Definition 4). Indeed, $\bar{X}$ and $\bar{Y}$ can be expressed as $\bar{\mu}_0$ and $\bar{\mu}_1$, respectively. The *global model*, i.e. the empirical mean over the union of a_X and a_Y's samples, is obtained by selecting $\alpha = n_Y / (n_X + n_Y)$.

The global model is optimal if X and Y have the same mean and variance, because it is the weight for which the variance of $\bar{\mu}_\alpha$ is minimal. However, a greater weight can be optimal (e.g., if $\mathbb{E}[Y] = \mathbb{E}[X]$ and $\text{Var}[Y] < \text{Var}[X]$), whereas a smaller weight is optimal if the true means of X and Y are very dissimilar or if $\text{Var}[Y] > \text{Var}[X]$.

To quantify the optimal weight α for any distributions X and Y, we express the ESE of $\bar{\mu}_\alpha$ as a function of α. We answer the questions: *Which weight $\alpha^\star$ would we tell a_X to use, if we had perfect knowledge of X and Y? Exactly how much smaller would the error $e_{\alpha^\star}$ be, relative to e_0? How large would the error e_α be if a_X used a different weight α?*

Definition 4 (Optimal model averaging weight). *We denote $\alpha^\star$ the weight which minimizes the ESE of $\bar{\mu}_{\alpha^\star}$ w.r.t. $\mathbb{E}[X]$:*

$$\alpha^\star \triangleq \arg\min_\alpha (e_\alpha)$$

3.3. Assumptions on X and Y

We make *no assumptions* on the data distributions X and Y — not even that they are related. However, our results in their strongest form depend on $\mathrm{Var}[X]$ being non-zero, and on X and Y having a finite mean and variance. Instead of making specific assumptions on X and Y, we express our results as functions of the unknown quantities $(\mathbb{E}[Y] - \mathbb{E}[X])^2$, $\mathrm{Var}[X]/n_X$, and $\mathrm{Var}[Y]/n_Y$. This approach allows us to quantify the potential of weighted model averaging exactly, for any distributions X and Y (where Y could be the union of all other agents' data sets in FL). On the other hand, since our results depend on unknown quantities, they cannot be used directly in empirical estimation problems. While it is certain that no empirical estimator based on the weighted averaging of $\bar{X}$ and $\bar{Y}$ can do better than the optimal estimator $\bar{\mu}_{\alpha^\star}$, it is much less clear that this bound can actually be attained in practice.

4. Theoretical Results

In Theorem 1, we find an analytic expression for the optimal model weight $\alpha^\star$. We conclude in Corollary 1.1, that the local empirical mean is dominated unless X is deterministic. Subsequently, we give the ESE of the weighted average $\bar{\mu}_\alpha$ for any α in Theorem 2, finding its minimum in Corollary 2.1 and its maximum in Corollary 2.4.

4.1. Prelude: Estimator Properties

In Lemmas 1 to 4, we calculate statistical properties of the estimators introduced in Section 3, including their ESE. The proofs of these basic results are relegated to Appendix A.

Lemma 1. *Let $\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_Y$, $\mathrm{Var}[X] = \sigma_X^2$, and $\mathrm{Var}[Y] = \sigma_Y^2$. We recall the properties of the empirical means $\bar{X}$ and $\bar{Y}$:*

$$\mathbb{E}[\bar{X}] = \mu_X, \qquad\qquad \mathbb{E}[\bar{Y}] = \mu_Y$$
$$\mathrm{Var}[\bar{X}] = \frac{\sigma_X^2}{n_X}, \qquad\qquad \mathrm{Var}[\bar{Y}] = \frac{\sigma_Y^2}{n_Y}$$

Lemma 2. *Recall the nomenclature of Lemma 1. The weighted average $\bar{\mu}_\alpha$ has the following properties:*

$$\mathbb{E}[\bar{\mu}_\alpha] = (1-\alpha)\,\mu_X \;\; + \alpha\mu_Y$$
$$\mathrm{Var}[\bar{\mu}_\alpha] = (1-\alpha)^2 \frac{\sigma_X^2}{n_X} + \alpha^2 \frac{\sigma_Y^2}{n_Y}$$

Lemma 3. *We compute the ESE of the empirical means:*

$$e_0 = \mathbb{E}\left[(\bar{X} - \mu_X)^2\right] = \frac{\sigma_X^2}{n_X}$$
$$e_1 = \mathbb{E}\left[(\bar{Y} - \mu_X)^2\right] = (\mu_Y - \mu_X)^2 + \frac{\sigma_Y^2}{n_Y}$$

Lemma 4. *We find the ESE of $\bar{\mu}_\alpha$ analytically below. It is convex w.r.t. α. If $\max\left\{(\mu_Y - \mu_X)^2, \sigma_X^2, \sigma_Y^2\right\} > 0$, then it is strictly convex.*

$$e_\alpha = (1-\alpha)^2 e_0 + \alpha^2 e_1$$

4.2. Optimal Model Averaging Weight

In Section 3.2, we defined the goal of agent a_X as selecting an optimal weight $\alpha^\star$ to minimize the ESE of $\bar{\mu}_\alpha$ w.r.t. μ_X. We prove the existence and (under minimal assumptions) uniqueness of $\alpha^\star$ in Theorem 1, where we also find its analytical form. This result is illustrated in Figure 2. In Corollary 1.1, we conclude that, unless $\mathrm{Var}[X] = 0$, the empirical mean $\bar{X}$ is always dominated by *some* weighted average $\bar{\mu}_\alpha$ with $\alpha > 0$. Finally, in Corollary 1.2 we prove an upper bound on $\alpha^\star$ that is simpler than Theorem 1.

Theorem 1. *Recall the setting and the estimators from Definitions 1, 2 and 4, as well as the nomenclature of Lemma 1. There **exists** an $\alpha^\star \in [0, 1]$ that minimizes the ESE of $\bar{\mu}_\alpha$ w.r.t. μ_X. Further, if $\max\left\{(\mu_Y - \mu_X)^2, \sigma_X^2, \sigma_Y^2\right\} > 0$, then $\alpha^\star$ is **unique** and satisfies the following equality:*

$$\alpha^\star = \frac{e_0}{e_0 + e_1}\left(= \frac{\mathrm{Var}[\bar{X}]}{\mathrm{Var}[\bar{X}] + (\mu_Y - \mu_X)^2 + \mathrm{Var}[\bar{Y}]}\right)$$

Proof of Theorem 1. Recall from Lemma 4 that the error is convex in α, from which we conclude that it admits at least one global minimum. **If** $(\mu_Y - \mu_X)^2 = 0 = \sigma_X^2 = \sigma_Y^2$, then the error vanishes irrespective of α (by Lemma 4). Thus, any α minimizes the error, including all $\alpha \in [0, 1]$. **Otherwise,** the error is strictly convex (by Lemma 4) and is uniquely minimized by $\alpha^\star$ s.t. $0 = \frac{\delta}{\delta\alpha}e_\alpha$. Thus, we find:

$$0 = \frac{\delta}{\delta\alpha}e_\alpha = 2(\alpha^\star - 1)e_0 + 2\alpha^\star e_1$$
$$\Leftrightarrow \quad 2e_0 = 2\alpha^\star(e_0 + e_1)$$
$$\overset{\max\{e_0, e_1\} > 0}{\Leftrightarrow} \alpha^\star = \frac{e_0}{e_0 + e_1}$$

In the last line above, we have made use of the assumption that $\max\left\{(\mu_Y - \mu_X)^2, \sigma_X^2, \sigma_Y^2\right\} > 0$, which implies that e_0 or e_1 is strictly positive.

To prove that $\alpha^\star \in [0, 1]$, we rewrite the last equation to obtain (using the non-negativity of the ESE):

$$\alpha^\star = \begin{cases} 0, & \text{if } e_0 = 0 \\ 0 \leq \frac{1}{1 + e_1/e_0} \leq 1, & \text{otherwise} \end{cases} \qquad \blacksquare$$

Using Theorem 1, we can ask exactly when the empirical mean $\bar{\mu}_0 = \bar{X}$ is optimal, leading to the following corollary:

Corollary 1.1. *Let the means and variances of X and Y be finite, and let $\sigma_X^2 > 0$. Then, the empirical mean $\bar{\mu}_0 = \bar{X}$ is dominated by some $\bar{\mu}_{\alpha^\star}$ with $\alpha^\star > 0$. Formally:*

$$\left(\sigma_X^2 \neq 0\right) \text{ and } \left(\max\{e_0, e_1\} < \infty\right) \implies \alpha^\star > 0$$

Proof of Corollary 1.1. By Lemma 3, e_0 and e_1 are finite if the means and variances of X and Y are finite. This, in turn, implies that the denominator of $\alpha^\star$ is bounded. Additionally, e_0 is strictly positive if $\sigma_X^2 \neq 0$. Hence, we conclude from Theorem 1 that $\alpha^\star$ is strictly positive under the stated conditions. ∎

Finally, we can prove simpler bounds on $\alpha^\star$ to improve our intuition: $\alpha^\star$ is only high when the local model has a lot of variance to trade away ($\text{Var}\left[\bar{X}\right]$ is large) and when $\bar{Y}$ has a low variance and is not unreasonably biased ($\text{Var}\left[\bar{Y}\right]$ and $(\mu_Y - \mu_X)^2$ are small):

Corollary 1.2. *Let $\sigma_X^2 > 0$. Then:*

$$\alpha^\star < \min\left\{\frac{\text{Var}\left[\bar{X}\right]}{(\mu_Y - \mu_X)^2}, \frac{\text{Var}\left[\bar{X}\right]}{\text{Var}\left[\bar{Y}\right]}\right\} \tag{1}$$

Proof of Corollary 1.2. $\sigma_X^2 > 0$ implies $\text{Var}\left[\bar{X}\right] > 0$, thus we write Theorem 1 as:

$$\alpha^\star = \frac{1}{1 + \frac{(\mu_Y - \mu_X)^2}{\text{Var}\left[\bar{X}\right]} + \frac{\text{Var}\left[\bar{Y}\right]}{\text{Var}\left[\bar{X}\right]}} < \frac{1}{\frac{(\mu_Y - \mu_X)^2}{\text{Var}\left[\bar{X}\right]} + \frac{\text{Var}\left[\bar{Y}\right]}{\text{Var}\left[\bar{X}\right]}} \tag{2}$$

The result follows by observing that $\text{Var}\left[\bar{X}\right]$, $(\mu_Y - \mu_X)^2$, and $\text{Var}\left[\bar{Y}\right]$ are non-negative. By abuse of notation, we allow division by 0 in Equation (1) and equate it to $+\infty$. ∎

4.3. Expected Squared Error

Armed with a formula for $\alpha^\star$, we now investigate the effect of (optimally) weighted model averaging on the ESE. Theorem 2, which is illustrated in Figure 1 (Section 5.1), relates the error of any $\bar{\mu}_\alpha$ to that of $\bar{X}$. In Corollaries 2.1 to 2.3, we calculate the ESE of $\bar{\mu}_{\alpha^\star}$, $\bar{\mu}_{2\alpha^\star}$, and $\bar{Y}$. Finally, we prove in Corollary 2.4 that the error of $\bar{\mu}_{\alpha^\star}$ is bounded by e_0 and e_1.

Theorem 2. *If $\alpha^\star$ is unique and positive (i.e., if $\sigma_X^2 > 0$), then we can express e_α as:*

$$e_\alpha = \left(1 + \alpha\left(\frac{\alpha}{\alpha^\star} - 2\right)\right)e_0$$

Proof of Theorem 2. We use our assumption ($\sigma_X^2 > 0 \implies e_0 > 0$) to express e_1 as a function of e_0:

$$e_1 = \frac{e_0}{e_0}\frac{(e_0 + e_1)\,e_1}{e_0 + e_1} = e_0\frac{e_0 + e_1}{e_0}\left(1 - \frac{e_0}{e_0 + e_1}\right)$$
$$= e_0\frac{1 - \alpha^\star}{\alpha^\star} = e_0\left(\frac{1}{\alpha^\star} - 1\right)$$

Inserting into Lemma 4 yields:

$$e_\alpha = (1-\alpha)^2 e_0 + \alpha^2 e_0\left(\frac{1}{\alpha^\star} - 1\right)$$
$$= e_0\left(1 - 2\alpha + \alpha^2 + \alpha^2\left(\frac{1}{\alpha^\star} - 1\right)\right)$$
$$= e_0\left(1 + \alpha\left(-2 + \frac{\alpha}{\alpha^\star}\right)\right) \quad ∎$$

Inserting $\alpha^\star$ into Theorem 2, we obtain the minimum error of weighted model averaging:

Corollary 2.1. *Estimating μ_X by $\bar{\mu}_{\alpha^\star}$ instead of $\bar{X}$ leads to a $(100 \cdot \alpha^\star)\%$ reduction in ESE if $\sigma_X^2 > 0$. Formally:*

$$e_{\alpha^\star} = (1 - \alpha^\star)\,e_0 = \min_{\alpha \in [0,1]} e_\alpha$$

Proof of Corollary 2.1. The result follows from Theorem 2, by observing that $\left(\frac{\alpha^\star}{\alpha^\star} - 2\right) = -1$. The equality $\min_{\alpha \in [0,1]} e_\alpha = e_{\alpha^\star}$ follows from Theorem 1. ∎

We obtain $e_{2\alpha^\star}$ and e_1 analogously by evaluating Theorem 2. Corollary 2.2 shows that $\alpha = 2\alpha^\star$ is no better than using $\bar{X}$, while Corollary 2.3 shows that the largest possible weight $\alpha = 1$ does better than $\bar{X}$ if $\alpha^\star > \frac{1}{2}$, and worse if $\alpha^\star < \frac{1}{2}$.

Corollary 2.2. *If $\sigma_X^2 > 0$, then the ESE of $\bar{\mu}_{2\alpha^\star}$ (w.r.t. μ_X) is equal to that of $\bar{X}$. Formally:*

$$e_{2\alpha^\star} = e_0$$

Corollary 2.3. *If $\sigma_X^2 > 0$, then the ESE of $\bar{Y}$ is given by:*

$$e_1 = \left(\frac{1}{\alpha^\star} - 1\right)e_0$$

As a counterpart to the global lower bound on $\alpha^\star$ found in Corollary 2.1, we prove a global upper bound below:

Corollary 2.4. *The ESE of $\bar{\mu}_\alpha$ is bounded above by:*

$$\max_{\alpha \in [0,1]} e_\alpha = \begin{cases} e_0, & \text{if } \sigma_X^2 > 0 \text{ and } \alpha^\star \geq \frac{1}{2} \\ e_1, & \text{otherwise} \end{cases}$$

Proof of Corollary 2.4. By its convexity (Lemma 4), e_α admits a maximum on the closed interval $[0,1]$, specifically at one of the extreme points of the interval. Thus, we need only compare e_0 and e_1. **If $\sigma_X^2 = 0$,** then $e_0 = 0 \leq e_1$. This implies that $\max_{\alpha \in [0,1]} e_\alpha = e_1$. **Otherwise,** we use Corollary 2.3 to find:

$$e_0 \geq e_1 \iff 1 \geq \frac{1}{\alpha^\star} - 1 \iff 2\alpha^\star \geq 1 \quad ∎$$

In summary, Theorem 1 and Corollary 1.1 show that there exists an optimal model averaging weight $\alpha^\star$ and that it is unique and positive. The effects of model averaging on the ESE are quantified in Theorem 2 and its corollaries.

5. Implications of our Findings

Throughout this section, we make the assumption that $\sigma_X^2 > 0$. Indeed, if X had zero variance, the local empirical mean $\bar{X}$ would neither need, nor permit, any further improvement.

We summarize our main results about the ESE of weighted model averaging in Section 5.1, and those about the optimal weight of $\bar{Y}$ in Section 5.2. Then, we present numerical examples in Section 5.3 to illustrate the various dependencies of $\alpha^\star$. In Section 5.4, we show that model averaging generalizes *shrinkage* estimators. Finally, in Section 5.5, we discuss the similarities and differences between our results and those obtained by Donahue & Kleinberg (2020).

5.1. Quality of the Weighted Model Average

We compute the ESE of the weighted average $\bar{\mu}_\alpha$, for any distributions X and Y and for any weight α, in Lemma 4. In Theorem 2, we express it as a function of the weight α, the optimal weight $\alpha^\star$, and e_0 (the error of the local empirical mean $\bar{X}$). This dependency is illustrated in Figure 1.
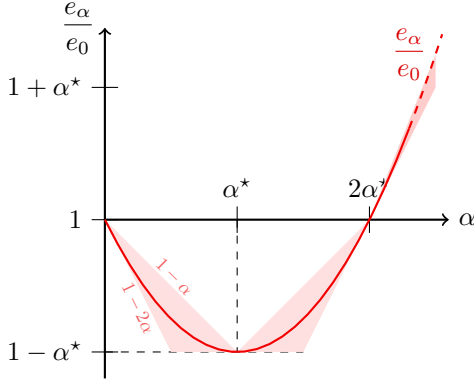

Figure 1. Illustration of Theorem 2. In red, we plot $\frac{e_\alpha}{e_0}$ depending on α and $\alpha^\star$. The shading represents simple linear bounds. The left bounds are annotated with the corresponding formulae, and the right shape is symmetrical. e_α: ESE of the weighted average $\bar{\mu}_\alpha$. e_0: ESE of the local estimator $\bar{X}$.

In Corollary 2.1, we conclude that $\bar{\mu}_\alpha$ can have an ESE lower than that of $\bar{X}$ by up to a fraction $\alpha^\star$, and no lower. However, the optimal weight $\alpha^\star$ is a function of the (unknown) means and variances of X and Y. It is unclear whether it is possible to achieve the same ESE with a purely empirical model average, whose weight does not depend on unknown distribution parameters.

Nevertheless, even a sub-optimally weighted model average can significantly improve upon the local model $\bar{X}$ if $\alpha^\star$ is not too small. Indeed, Figure 1 illustrates that using $\bar{\mu}_\alpha$ with $\alpha < \alpha^\star$ reduces the error by more than a fraction of α, and that any $0 < \alpha < 2\alpha^\star$ has $e_\alpha < e_0$. Specifically, all $0 < \alpha < 1$ yield an improvement over $\alpha = 0$ if $\mathrm{Var}\big[\bar{X}\big] \geq$
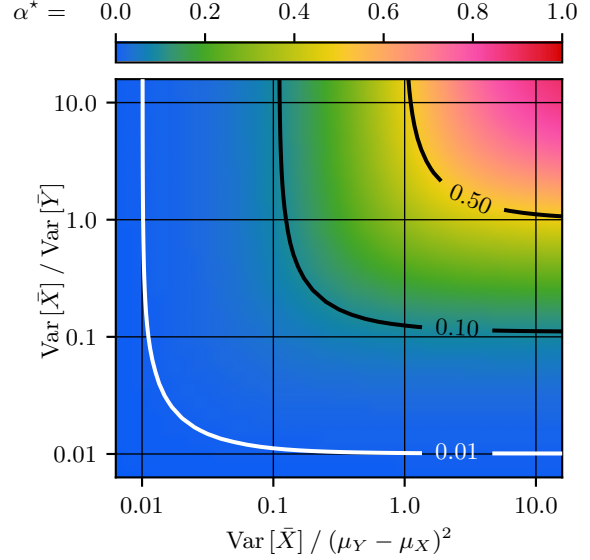

Figure 2. Illustration of Theorem 1, showing the value of $\alpha^\star$ depending on $\frac{\mathrm{Var}[\bar{X}]}{(\mu_Y - \mu_X)^2}$ and $\frac{\mathrm{Var}[\bar{X}]}{\mathrm{Var}[\bar{Y}]}$ (cf. Corollary 1.2). The value of $\alpha^\star$ is mapped to a color scale, and contour lines are drawn at $\alpha^\star = 0.01$, $\alpha^\star = 0.1$, and $\alpha^\star = 0.5$.

$(\mu_Y - \mu_X)^2 + \mathrm{Var}\big[\bar{Y}\big]$ (because this implies $2\alpha^\star \geq 1$).

What Figure 1 cannot show, however, is the tight upper bound to e_α proven in Corollary 2.4. For $\alpha^\star \geq \frac{1}{2}$, e_α is upper-bounded by e_0. Otherwise, it is upper-bounded by e_1, which is related to e_0 and $\alpha^\star$ by Corollary 2.3. Thus, this upper bound can be added to the figure for a given $\alpha^\star$ by drawing a vertical line at $\alpha = 1$. If $\alpha^\star \geq \frac{1}{2}$, then the vertical line crosses $\frac{e_\alpha}{e_0}$ at a value below 1. Corollary 2.3 shows that the error can be increased substantially over that of $\bar{X}$, especially if $\alpha^\star$ is very close to 0.

5.2. Optimal Model Averaging Weight

In Section 5.1, we show that $\alpha^\star$ is an upper bound to the potential benefits of weighted model averaging, and that a small value of $\alpha^\star$ opens up the possibility of doing worse than $\bar{X}$. This central role motivates us to visualize $\alpha^\star$ and its dependency on the relevant parameters. In Figure 2, the contour lines help appreciate Corollary 1.2, which states that $\alpha^\star$ is bounded above by the lower of the two terms: $\mathrm{Var}\big[\bar{X}\big] / (\mu_Y - \mu_X)^2$ (on the horizontal axis), and $\mathrm{Var}\big[\bar{X}\big] / \mathrm{Var}\big[\bar{Y}\big]$ (on the vertical axis).

Indeed, the contour lines are approximately horizontal when $\frac{\mathrm{Var}[\bar{Y}]}{(\mu_Y - \mu_X)^2} > 5$ (bottom right corner of Figure 2), and vertical when $\frac{\mathrm{Var}[\bar{Y}]}{(\mu_Y - \mu_X)^2} < \frac{1}{5}$ (top left corner). What's more, the contour lines for $\alpha^\star = 0.01$ and $\alpha^\star = 0.1$ approximately asymptotically approach the 0.01 and 0.1-grid-lines respec-

tively, as the distance from the diagonal increases. Thus, the bound in Corollary 1.2 is approximately tight as the relative difference between $\mathrm{Var}[\bar{Y}]$ and $(\mu_Y - \mu_X)^2$ increases. This approximation is particularly good if the bound is much smaller than 1, and falls apart of course when it is greater than 1—as exemplified by the contour line at $\alpha^\star = 0.5$.

As a consequence of this bound, increasing n_Y yields diminishing returns: Doubling n_Y almost doubles $\alpha^\star$ when both $\mathrm{Var}[\bar{X}]$ and $(\mu_Y - \mu_X)^2$ are much smaller than $\mathrm{Var}[\bar{Y}]$, but $\alpha^\star$ is barely affected if one of these terms is much greater than $\mathrm{Var}[\bar{Y}]$. If the variance of $\bar{X}$ is much greater than that of $\bar{Y}$, however, $\alpha^\star$ may already be close to 1. In this case, a further increase of n_Y may no longer produce a big change in $\alpha^\star$, but it may still reduce $e_{\alpha^\star}$ quite drastically by lowering the term $(1 - \alpha^\star)$.

Importantly, both terms in Corollary 1.2 feature the variance of $\bar{X}$ (which equals e_0). Naturally, as the local estimator grows more accurate with increasing n_X, the need for collaboration vanishes along with its potential benefits. Moreover, even for small n_X, model averaging is useful only if σ_X^2 is not too small compared to $(\mu_Y - \mu_X)^2$ and σ_Y^2.

5.3. Numerical Examples

To give some intuition of its dependencies, we calculate $\alpha^\star$ for a variety of scenarios in Table 1. We also use the value of $\alpha^\star$ to calculate the relative size of the error for three estimators, compared to the error of $\bar{X}$:

$\bar{\mu}_{\alpha^\star}$ the optimally weighted model average,
$\bar{\mu}_{1/5}$ the average with a weight of 20% (i.e., $\frac{4}{5}\bar{X} + \frac{1}{5}\bar{Y}$),
$\bar{\mu}_{1/2}$ the average with a weight of 50% (i.e., $\frac{1}{2}\bar{X} + \frac{1}{2}\bar{Y}$).

We describe each scenario by four easily interpretable quantities (Table 1, left of the vertical bar):

$\frac{(\mu_Y - \mu_X)^2}{\sigma_X^2}$ the squared bias of Y w.r.t. X, relative to $\mathrm{Var}[X]$,
n_X the number of samples drawn from X,
$\frac{\sigma_Y^2}{\sigma_X^2}$ the variance of Y, relative to that of X,
$\frac{n_Y}{n_X}$ the number of samples drawn from Y, relative to n_X.

However, several configurations of these four variables are equivalent because $\alpha^\star$ is fully determined by the terms $\frac{(\mu_Y - \mu_X)^2}{\sigma_X^2} \cdot n_X$ and $\left(\frac{\sigma_Y^2}{\sigma_X^2}\right) / \left(\frac{n_Y}{n_X}\right)$ (cf. Equation (2)). This is why the second, fourth, sixth, and twelfth ($n_X = 5$) rows of Table 1 are equivalent to their respective predecessor.

The second row demonstrates that setting $n_Y \to \infty$ is equivalent to setting $\sigma_Y^2 \to 0$. The fourth and sixth row show that the multiplication of σ_Y^2 with a constant can be compensated by multiplying n_Y with the same constant. Furthermore, the third row of Table 1 embodies the diminishing returns phenomenon described in Section 5.2: raising n_Y until

$\left(\frac{n_Y}{n_X}\right) = 6\left(\frac{\sigma_Y^2}{\sigma_X^2}\right)$ already takes $\alpha^\star$ most of the way towards its limit for $n_Y \to \infty$. Finally, the eleventh and twelfth rows (from $n_X = 20$ to $n_X = 5$) exemplify the connection between the first two columns, which is that $\alpha^\star$ only depends on their product. This also explains the equivalence of the last two rows of Table 1. It also reflects Corollary 1.2, as the influence of n_Y is largest when $\frac{(\mu_Y - \mu_X)^2}{\sigma_X^2} \cdot n_X$ is small, and becomes vanishingly small when the product exceeds ~ 10 (see bottom rows). Another interpretation is that, when $\mu_Y = \mu_X \pm \sigma_X$ (i.e., $(\mu_Y - \mu_X)^2 = \sigma_X^2$), then using μ_Y as a proxy for μ_X is exactly as good as using a single realization of X (in terms of ESE).[2]

Table 1. Numerical examples for the value of $\alpha^\star$ in a variety of scenarios. The corresponding errors of $\bar{\mu}_{\alpha^\star}$, $\bar{\mu}_{1/5}$, and $\bar{\mu}_{1/2}$ are given as fractions (or multiples) of e_0. Cells marked with $*$ could contain any finite positive value without affecting $\alpha^\star$, while empty cells have the same value as the cell above. bias^2 denotes $(\mu_Y - \mu_X)^2$.

$\frac{\mathrm{bias}^2}{\sigma_X^2}$	n_X	$\frac{\sigma_Y^2}{\sigma_X^2}$	$\frac{n_Y}{n_X}$	$\alpha^\star$	$\frac{e_{\alpha^\star}}{e_0}$	$\frac{e_{1/5}}{e_0}$	$\frac{e_{1/2}}{e_0}$
0	$*$	0	$*$	**1.0**	0.00	0.64	0.25
		$*$	$+\infty$	**1.0**			
		1	6	**0.86**	0.14	0.65	0.29
		10	60	**0.86**			
		1	1	**0.50**	0.50	0.68	0.50
		10	10	**0.50**			
0.25	10	0	$+\infty$	**0.57**	0.43	0.67	0.44
		1	1	**0.44**	0.56	0.69	0.56
	100	0	$+\infty$	**0.04**	0.96	1.64	6.50
		1	1	**0.04**	0.96	1.68	6.75
	20	0	$+\infty$	**0.17**	0.83	0.84	1.50
1	5	0	$+\infty$	**0.17**			
		1	1	**0.14**	0.86	0.88	1.75
	50	0	$+\infty$	**0.02**	0.98	2.64	12.8
		1	1	**0.02**	0.98	2.65	13.0
$*$	$+\infty$	$*$	$*$	**0.0**	1.00	$+\infty$	$+\infty$
$+\infty$	$*$	$*$	$*$	**0.0**			

5.4. Relation to the James-Stein Estimator

Weighted model averaging generalizes the fundamental statistical notion of *shrinkage*. Conceptually, *shrinkage* (towards 0) simply corresponds to multiplying an empirical estimator with a weight $\beta < 1$. This reduces the estimator's variance at the cost of additional bias, thus reducing its ESE if β is sufficiently close to 1 (Gruber, 2017). For $\sigma_Y^2 = 0$, we recover the case of estimator shrinkage towards an arbitrary anchor value (μ_Y). If we pick $Y = 0$, then $\bar{\mu}_\alpha$ is simply a shrunken version of $\bar{X}$: $\bar{\mu}_\alpha = (1 - \alpha)\bar{X}$.

We find that the optimal amount of shrinkage is never *quite* 0

---

[2]   Indeed, after drawing one sample $x_1 \sim X$, we find: $\mathbb{E}\left[(\mu_Y - \mu_X)^2\right] = (\mu_Y - \mu_X)^2 = \sigma_X^2 = \mathbb{E}\left[(x_1 - \mu_X)^2\right]$

for any combination of finite means, variances, and number of samples. However, we see in Table 1 that it quickly tends to 0 as we shrink towards an increasingly unrelated value (with $\frac{(\mu_Y - \mu_X)^2}{\sigma_X^2} \rightarrow \infty$), and as the quality of the local estimator $\bar{X}$ increases (with $n_X \rightarrow \infty$).

The first example of shrinkage is the infamous James-Stein (JS) estimator, which dominates the empirical mean for the problem of estimating from just one observation the mean of a spherically symmetrical multivariate normal random variable (RV) with three or more dimensions (Gruber, 2017; James & Stein, 1961). While the JS estimator can be used without prior knowledge of the mean and variance of the RV, it only dominates the empirical mean under the assumptions that all coordinates of the RV (a) have the same variance, and that they are (b) normal and (c) mutually uncorrelated. It is also restricted to estimation in at least three dimensions. By contrast, our results hold for arbitrary one-dimensional estimation problems, with minimal assumptions on X and Y. However, we express them as functions of statistics of X and Y that would not be available in a practical setting.

5.5. Connection to Recent Related Work

Donahue & Kleinberg (2020) prove corresponding results to our Theorem 1 and Corollaries 1.1 and 2.1 in their Lemmas 6.1 and 6.3, under comparatively more complex assumptions. Our model is simpler in three ways, making it more general for some aspects and more specific for others.

The most important difference is that our assumptions on the input data are minimal and significantly more realistic, as we assume nothing more than finite means and variances, and a non-zero variance for X. In contrast, Donahue & Kleinberg (2020) assume that μ_Y and μ_X are drawn independently from the same random variable with variance σ^2, and that X and Y have the same variance ($\sigma_X^2 = \sigma_Y^2 = \mu_e$).

Secondly, we consider only two nodes ($M = 2$), whereas their analysis includes an arbitrary number of players to study clustering approaches to model personalization. Nevertheless, our two-node model reproduces their *coarse-grained federation* model (Section 6 in their paper) by setting: $\bar{Y} = \frac{1}{N} \sum_{i=1}^{M} \hat{\theta}_i \cdot n_i$. Finally, our analysis is confined to one-dimensional linear regression ($D = 1$) compared to theirs in D dimensions.

We will demonstrate that these corresponding results become equivalent in the special case when both respective classes of assumptions are applied jointly.

Lemma 5. *By applying the assumptions of Donahue & Kleinberg (2020) on the data generation process, to Corollary 2.1, we obtain the same result as by applying our assumption of $M = 2$ nodes to the one-dimensional ($D = 1$) version of their Lemma 6.3.*

Lemma 5 follows from Lemmas 5.1 and 5.2. The proof of Lemma 5.1 is relegated to Appendix B.

Lemma 5.1. *Under the assumptions that (i) μ_Y and μ_X are drawn independently from the same random variable W with variance σ^2 and (ii) $\sigma_X^2 = \sigma_Y^2 = \mu_e$, Corollary 2.1 and Theorem 1 yield:*

$$e_{\alpha^\star} = \frac{2n_Y^2 \sigma^2 + \mu_e n_Y}{n_Y (n_Y + n_X) + 2n_X n_Y^2 \frac{\sigma^2}{\mu_e}}$$

Lemma 5.2. *Under the assumption that $M = 2$, denoting n_j by n_X and $n_{i \neq j}$ by n_Y, Lemma 6.3 from Donahue & Kleinberg (2020) simplifies to:*

$$MSE = \frac{2n_Y^2 \sigma^2 + \mu_e n_Y}{n_Y (n_Y + n_X) + 2n_X n_Y^2 \frac{\sigma^2}{\mu_e}}$$

Proof. Lemma 5.2 is obtained with the substitutions:

$$n_j = n_X \qquad\qquad N = n_X + n_Y$$
$$\sum_{i \neq j} n_i^2 = n_Y^2 \qquad\qquad N - n_j = n_Y \qquad \blacksquare$$

6. Conclusion

In this work, we have quantified the improvement of quality (reduction in expected squared error) that can be achieved through weighted model averaging with peers, as compared to using exclusively local data. Our results concern the mean estimation problem for any scalar random variable X, and include the case of shrinkage towards a fixed value. While we limit our analysis to averaging between two models $\bar{X}$ and $\bar{Y}$, our results apply to federated learning by letting a_Y denote the union of all other agents. We derive the optimal averaging weight for the helper model $\bar{Y}$, proving that it is strictly positive unless $\mathrm{Var}[X] = 0$. Through examples, we explain that this holds true even for an arbitrarily unsuitable helper model $\bar{Y}$ (thus for arbitrary non-IID client data), though the optimal weight converges towards 0 as the difference between the expectations of X and Y increases.

Further, we show that model averaging with the optimal weight $\alpha^\star$ reduces the ESE by a fraction equal to $\alpha^\star$, compared to the local model $\bar{X}$. We then analyse the ESE of the weighted model average, as a function of only the optimal weight $\alpha^\star$ and the weight α that is actually used. We find that any weight $\alpha < 2\alpha^\star$ reduces the ESE compared to $\bar{X}$, by more than a fraction of α if $\alpha < \alpha^\star$, and that any weight of $\alpha > 2\alpha^\star$ results in a larger ESE, compared to $\bar{X}$.

Motivated by its central role, we visualize and investigate the dependencies of the optimal model averaging weight $\alpha^\star$. It depends mainly on the lower of two ratios, which both involve the variance of the local estimator, $\mathrm{Var}[\bar{X}]$. Thus, $\alpha^\star$ can only be large if $\mathrm{Var}[\bar{X}]$ is significantly larger than the variance of the helper model $\bar{Y}$ and the squared difference between the expectations of X and Y.

Outlook. We prove theorems under realistic assumptions about the distributions of X and Y, in the limited setting of one-dimensional parameter estimation with two agents. Future work could extend our results to multivariate parameter estimation and to an arbitrary number of agents. Further, in addition to interpreting it *as* shrinkage, we could apply shrinkage *to* the weighted model average, thus obtaining $\mu_{\beta,\gamma} = (1 - \beta)\,\bar{X} + \gamma\bar{Y}$, where $\beta \geq \gamma$.

Finally, whether our theoretical lower bound on the ESE of the weighted model average can possibly be achieved without perfect knowledge of the bias and variances of X and Y, merits investigation. It would thus be interesting to quantify the error reduction that can be achieved in practice, potentially under specific assumptions on X and Y.

References

Deng, Y., Kamani, M. M., and Mahdavi, M. Adaptive Personalized Federated Learning. *arXiv:2003.13461 [cs, stat]*, March 2020. URL http://arxiv.org/abs/2003.13461.

Donahue, K. and Kleinberg, J. Model-sharing games: Analyzing federated learning under voluntary participation. *arXiv preprint arXiv:2010.00753*, 2020. URL http://arxiv.org/abs/2010.00753.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized federated learning: A meta-learning approach. *arXiv preprint arXiv:2002.07948*, 2020.

Grimberg, F., Hartley, M.-A., Jaggi, M., and Karimireddy, S. P. Weight erosion: An update aggregation scheme for personalized collaborative machine learning. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pp. 160–169. Springer, 2020.

Gruber, M. *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*. Routledge, 2017.

James, W. and Stein, C. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 361–379, Berkeley, Calif., 1961. University of California Press. URL https://projecteuclid.org/euclid.bsmsp/1200512173.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., and Zhao, S. Advances and Open Problems in Federated Learning. *arXiv:1912.04977 [cs, stat]*, December 2019. URL http://arxiv.org/abs/1912.04977.

Konečný, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *arXiv:1610.02527 [cs]*, October 2016. URL http://arxiv.org/abs/1610.02527.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Processing Magazine*, 37(3):50–60, May 2020. ISSN 1558-0792. doi: 10.1109/MSP.2020. 2975749. Conference Name: IEEE Signal Processing Magazine.

Lin, T., Kong, L., Stich, S. U., and Jaggi, M. Ensemble distillation for robust model fusion in federated learning. *NeurIPS - Advances in Neural Information Processing Systems*, 33, 2020.

Mansour, Y., Mohri, M., Ro, J., and Suresh, A. T. Three Approaches for Personalization with Applications to Federated Learning. *arXiv:2002.10619 [cs, stat]*, February 2020. URL http://arxiv.org/abs/2002. 10619.

Nedic, A. Distributed gradient methods for convex machine learning problems in networks: Distributed optimization. *IEEE Signal Processing Magazine*, 37(3):92–101, 2020.

Singh, S. P. and Jaggi, M. Model fusion via optimal transport. *NeurIPS - Advances in Neural Information Processing Systems*, 33, 2020.

Su, Y., Dimakopoulou, M., Krishnamurthy, A., and Dudik, M. Doubly robust off-policy evaluation with shrinkage. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9167–9176, Virtual, 13–18 Jul 2020. PMLR. URL http://proceedings.mlr.press/v119/ su20a.html.

Wang, K., Mathews, R., Kiddon, C., Eichner, H., Beaufays, F., and Ramage, D. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019. URL http://arxiv.org/abs/1910.10252.

Optimal Weighted Model Averaging for Personalized Collaborative Learning
Appendix: Lemma Proofs

A. Lemmas in Section 4

Proof of Lemma 1. The proofs for $\bar{X}$ apply analogously to $\bar{Y}$. The unbiasedness of the empirical mean ($\mathbb{E}[\bar{X}] = \mathbb{E}[X]$) follows from the linearity property of the expectation. The variance of $\bar{X}$ follows from the independence of the realizations $\{x_1, \ldots, x_{n_X}\}$: ∎

Proof of Lemma 2. The expectation of $\bar{\mu}_\alpha$ follows from the linearity property of the expectation. Its variance also results from the properties of the variance:

$$\text{Var}[\bar{\mu}_\alpha] = \text{Var}\big[(1-\alpha)\bar{X} + \alpha\bar{Y}\big]$$
$$= (1-\alpha)^2\,\text{Var}[\bar{X}] + \alpha^2\,\text{Var}[\bar{Y}]$$

The result follows from Lemma 1. ∎

Proof of Lemma 3. To compute the ESE of the empirical means, we make use of the bias-variance decomposition:

$$\mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right] = \mathbb{E}[\hat{\theta}^2 - 2\theta\hat{\theta} + \theta^2]$$

$$= \underbrace{\mathbb{E}[\hat{\theta}^2] - \left(\mathbb{E}[\hat{\theta}]\right)^2}_{} \overbrace{+ \left(\mathbb{E}[\hat{\theta}]\right)^2}^{\pm 0} \underbrace{- 2\theta\mathbb{E}[\hat{\theta}] + \theta^2}_{}$$

$$= Var[\hat{\theta}] + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2 \qquad (3)$$

Plugging in μ_X, $\bar{X}$ and $\bar{Y}$, we obtain:

$$\mathbb{E}\left[\left(\bar{X} - \mu_X\right)^2\right] = \text{Var}[\bar{X}] + \overbrace{\left(\mathbb{E}[\bar{X}] - \mu_X\right)^2}^{0}$$
$$\mathbb{E}\left[\left(\bar{Y} - \mu_X\right)^2\right] = \text{Var}[\bar{Y}] + \left(\mathbb{E}[\bar{Y}] - \mu_X\right)^2$$

The result follows from Lemma 1. ∎

Proof of Lemma 4. We derive our analytic formula for $\alpha^\star$ by plugging the results from Lemma 2 into the bias-variance decomposition (Equation (3)):

$$e_\alpha = \mathbb{E}\left[(\bar{\mu}_\alpha - \mu_X)^2\right] = \text{Var}[\bar{\mu}_\alpha] + (\mathbb{E}[\bar{\mu}_\alpha] - \mu_X)^2$$
$$= (1-\alpha)^2 \frac{\sigma_X^2}{n_X} + \alpha^2 \frac{\sigma_Y^2}{n_Y} + (\alpha(\mu_Y - \mu_X))^2$$

Using Lemma 3, we obtain:

$$e_\alpha = (1-\alpha)^2 e_0 + \alpha^2 e_1$$

We prove the convexity results by taking the second derivative of e_α w.r.t. α. We obtain the first derivative trivially as:

$$\frac{\delta}{\delta\alpha} e_\alpha = 2(\alpha - 1)e_0 + 2\alpha e_1$$

The second derivative of the error w.r.t. α follows as:

$$\frac{\delta^2}{\delta\alpha^2} e_\alpha = 2(e_0 + e_1) \geq 0$$

We conclude the proof by observing that the second derivative of the error w.r.t. α is a sum of non-negative terms. Thus, it is itself non-negative. By Lemma 3, it is positive if $\max\left\{(\mu_Y - \mu_X)^2, \sigma_X^2, \sigma_Y^2\right\} > 0$. ∎

B. Lemmas in Section 5

Proof of Lemma 5.1. Plugging Theorem 1 into Corollary 2.1 yields:

$$e_{\alpha^\star} = (1 - \alpha^\star)e_0 = \frac{e_1}{e_0 + e_1} e_0$$

Using Lemma 3, we obtain:

$$e_{\alpha^\star} = \frac{\mathbb{E}[(\mu_Y - \mu_X)^2] + \frac{\sigma_Y^2}{n_Y}}{\frac{\sigma_X^2}{n_X} + \mathbb{E}[(\mu_Y - \mu_X)^2] + \frac{\sigma_Y^2}{n_Y}} \frac{\sigma_X^2}{n_X}$$

Using assumption (ii):

$$e_{\alpha^\star} = \frac{\mathbb{E}[(\mu_Y - \mu_X)^2] + \frac{\mu_e}{n_Y}}{\frac{\mu_e}{n_X} + \mathbb{E}[(\mu_Y - \mu_X)^2] + \frac{\mu_e}{n_Y}} \frac{\mu_e}{n_X} \qquad (4)$$

Under assumption (i), we observe:

$$\mathbb{E}\left[(\mu_Y - \mu_X)^2\right] = \text{Var}[(\mu_Y - \mu_X)] + \underbrace{(\mathbb{E}[\mu_Y - \mu_X])^2}_{=0}$$
$$= \text{Var}[\mu_Y] + \text{Var}[\mu_X] = 2\,\text{Var}[W] = 2\sigma^2$$

Substituting in Equation (4) yields:

$$e_{\alpha^\star} = \frac{2\sigma^2 + \frac{\mu_e}{n_Y}}{\frac{\mu_e}{n_X} + 2\sigma^2 + \frac{\mu_e}{n_Y}} \frac{\mu_e}{n_X}$$
$$= \frac{2\sigma^2 + \frac{\mu_e}{n_Y}}{1 + 2n_X\frac{\sigma^2}{\mu_e} + \frac{n_X}{n_Y}}$$
$$= \frac{2n_Y^2\sigma^2 + \mu_e n_Y}{n_Y^2 + n_Y n_X + 2n_X n_Y^2 \frac{\sigma^2}{\mu_e}} \qquad ∎$$