Personalized heterogeneous Gaussian mean estimation under communication constraints

Ruida Zhou and Suhas Diggavi
Department of Electrical and Computer Engineering
University of California, Los Angeles
Email: ruida@ucla.edu and suhas@ee.ucla.edu

***Abstract*—We consider personalized estimation for heterogeneous data under communication constraints. In many applications, distributed users have heterogeneous local data with distinct statistics, and want to estimate individual (personalized) properties of the local data. However, they have limited local data and we explore how collaboration (even over communication-limited links) can enable better personalized estimation. We study this for the Gaussian Bayesian model for heterogeneity with unknown parameters and a worst-case total regret criterion. We characterize (order-wise) the worst-case regret for personalized mean estimation by devising novel lower bounds and achievability schemes, which also demonstrates the value of collaboration.**

I. Introduction

In many applications (*e.g.,* medical) the data of each client could have distinct characteristics (statistics), and naturally one wants to estimate parameters related to the individual data. However, there might not be enough local samples to obtain a good individual (personalized) estimate. Therefore a natural question is whether collaboration with other clients with different data statistics could help personalized estimation. In order to address this question sysematically, one needs to formulate the statistical question, and a framework advocated recently [5] is to use a hierarchical model to capture the data heterogeneity. In this, there is a population distribution from which local parameters are sampled, and in turn, local data is generated with these local parameters capturing statistical heterogeneity. It has been shown in [5] that this model captures the most practical methods for personalized estimation and learning, as well as suggests new ones.

In addition, in applications such as federated (distributed) learning, the local data resides in remote clients which have constrained communication links for collaboration. Therefore this leads to the question of the fundamental trade-off between communication constraints and personalized estimation. We formulate these questions through the criterion of a total regret formalization (see (1) for precise definition), which is the difference in loss between the estimator which has access to the population distribution and an estimator which does not have this knowledge, where nature gets to maximize regret by choosing the population distribution.

We focus on the case where the population distribution G is a vector Gaussian with unknown mean $\boldsymbol{\mu}$, and unknown co-variance $\sigma^2\mathbb{I}_d$. The local parameters $\{\boldsymbol{\theta}_i\}_{i=1}^m$ are generated from G and in turn generate local n data samples $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(n)}$ which are generated as Gaussian with mean $\boldsymbol{\theta}_i$ and known covariance Σ_o. The goal is to find a personalized estimator $\hat{\boldsymbol{\theta}}_i$, which can depend on the local data $\mathbf{x}_i^{(1:n)}$ and potentially collaborative messages $\{Z_j\}_{j\neq i}$ from the other clients. We consider the communication-constrained case where the messages $\{Z_j\}$ have a given entropy bound.

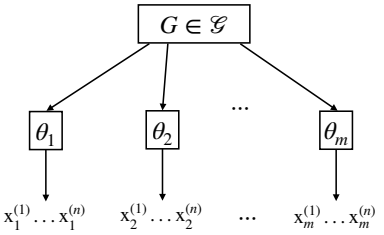

Fig. 1. An illustrative figure of the personalized estimation under a Bayesian model with m users and each user having n samples.

Contributions: We first examine the problem without communication constraints and develop both information-theoretic lower bounds and a scheme to completely characterize the total regret in Theorem III.3. Perhaps of interest is that the lower bound is developed through a hyper-prior technique, which gives a sharper bound than the more conventional Le Cam bounding technique. Next, we extend this to the case with communication constraints and characterize (order-wise) total regret in Theorem IV.1. In this case as well, we develop information-theoretic lower bounds (Theorem IV.4) as well as a scheme (Theorem IV.3) which uses a functional representation result introduced in [4].

Organization: We first formally state the problem and setup notations in Section II. We give the results without communication constraints in Section III and the results with communication constraints in Section IV. Many of the proof details are given in supplementary appendices.

II. Problem Setting

There are a total of m users, and each user-i aims to estimate an unknown vector $\boldsymbol{\theta}_i \in \mathbb{R}^d$ by local observations $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \ldots, \mathbf{x}_i^{(n)} \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{\theta}_i, \Sigma_o)$. We here consider the case of spherical Gaussian with covariance matrix $\Sigma_o = \sigma_o^2\mathbb{I}_d$.

We assume there is an *unknown population distribution (a.k.a. prior in Bayesian models)* G from a class of the distributions $\mathcal{G}$ over $\mathbb{R}^d$, such that $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_m \overset{i.i.d.}{\sim} G$. Once G is known, it becomes a canonical Bayesian model as illustrated in Fig. 1. Without the knowledge of G, it belongs to the empirical/hierarchical Bayes setting, and the performance of local tasks can still be improved by sharing information among the users. In this work, we consider a Gaussian population

distribution $G = \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbb{I}_d)$ with parameters $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\sigma^2 \geq 0$, which are both unknown.

The communication protocol is defined as follows. Each user-i encodes the local observations $\mathbf{x}_i^{(1:n)}$ into a message $Z_i = f_i(\mathbf{x}_i^{(1:n)}, U_i)$ according to some encoding function f_i and randomness U_i independent of data. The message is distributed/broadcast to other users. After receiving the messages $Z_{-i} \triangleq Z_{1:i-1,i+1:m}$, user-$i$ estimates the unknown vector of interests $\boldsymbol{\theta}_i$ by some estimator $\hat{\boldsymbol{\theta}}_i(\mathbf{x}_i^{(1:n)}, Z_{-i})$. We consider the case that $\sup_{G \in \mathcal{G}} H_G(Z_i) \leq B$, where $H_G(Z_i)$ is the entropy under the prior G.

A. Performance metric

We consider the mean square error (MSE). The MSE of an estimator $\hat{\boldsymbol{\theta}}$ of the unknown vector $\boldsymbol{\theta}$ is $L(\hat{\boldsymbol{\theta}}; G) = \mathbb{E}_G[\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2]$, where $\mathbb{E}_G$ indicates that the expectation is taken under the Bayesian model with prior G.

The performance of personalized distributed estimators is measured by the total regret

$$\begin{aligned}\text{TotReg}(\hat{\boldsymbol{\theta}}_{1:m}; G) &= \sum_{i=1}^m (L(\hat{\boldsymbol{\theta}}_i; G) - L(\hat{\boldsymbol{\theta}}_G; G)) \\ &\overset{(a)}{=} \sum_{i=1}^m \mathbb{E}_G \left[\|\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_G\|^2 \right],\end{aligned}$$

where $\hat{\boldsymbol{\theta}}_G = \mathbb{E}_G[\boldsymbol{\theta}_i | \mathbf{x}_i^{(1:n)}]$ is the Bayes optimal estimator under prior G, a.k.a. minimum MSE (MMSE) estimator. The total regret is non-negative and measures the loss due to not knowing the true population distribution G, as the first equation shows that it is the difference between the MSE of the estimator without the knowledge of G and the MSE of the MMSE estimator with the knowledge of G. Equation (a) is by the orthogonality of the optimal estimator that $\mathbb{E}[(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_G)^\top (\hat{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_i)] = 0$, and it interprets the regret as MSE w.r.t. the Bayes estimator $\hat{\boldsymbol{\theta}}_G$.

With a slight abuse of notation, we define the worst-case total regret and minimax total regret as

$$\begin{aligned}\text{TotReg}(\hat{\boldsymbol{\theta}}_{1:m}; \mathcal{G}) &= \sup_{G \in \mathcal{G}} \text{TotReg}(\hat{\boldsymbol{\theta}}_{1:m}; G), \\ \text{TotReg}(\mathcal{G}, B) &= \inf_{\hat{\boldsymbol{\theta}}_{1:m}} \sup_{G \in \mathcal{G}} \text{TotReg}(\hat{\boldsymbol{\theta}}_{1:m}; G). \qquad (1)\end{aligned}$$

The infimum is taken over the estimator (as well as the encoding functions $f_{1:m}$) that satisfies the required communication constraints of B. Without the communication constraints, i.e., $B = \infty$, we simply let $\text{TotReg}(\mathcal{G}) = \text{TotReg}(\mathcal{G}, \infty)$. We assume $m > 3$ and in the vector case $d > 3$.

B. Related works

The estimation with single one-dimensional local observation and without constraints, i.e., $n = 1$ and $Z_i = \mathbf{x}_i^{(1)} \in \mathbb{R}$, has been studied in the nonparametric setting, i.e., G can be any sub-Gaussian distribution, where the state-of-the-art results of a total regret of $O((\log m)^5)$ is achieved by the nonparametric maximum likelihood estimator (NPMLE) in [3] and a lower bound of $\Omega((\log m)^2)$ is achieved by Assouad's method [6]. The estimation with multiple observations without constraints has been studied in the Gaussian population distribution with known variance setting [5]. Our work considers more general d-dimensional observations in the Gaussian population distribution with both unknown mean and variance setting with limited communication.

The empirical Bayes method is powerful for large throughput data, i.e., m and d are large. Considering a hierarchical Bayesian model can indeed motivates the design of estimators, which will be illustrated in this paper, e.g., the proposed hierarchical James-Stein estimators in Eq (7) and Eq (8).

III. Personalized estimation without constraints

We first present new results for the classic Gaussian mean estimation under the empirical Bayes model, where the message Z_i from user-i can be its local data $\mathbf{x}_i^{(1:n)}$ without communication constraints, illustrating the effect of aggregating information among users and introduce estimators that are useful under communication constraints.

A. Scalar case

The parameter $\theta_i \in \mathbb{R}$ has a Gaussian population $G = \mathcal{N}(\mu, \sigma^2)$. We here simply consider one observation, e.g., $n = 1$, and hence change $\mathbf{x}_i^{(1)}$ to x_i for notational convenience. It is straightforward to generalize it to the setting with multiple local samples, since their sample average is a sufficient statistic for parameter θ_i. In this case, the MMSE estimator is

$$\hat{\theta}_G(x) = \mathbb{E}_G[\theta | \mathrm{x} = x] = x + \frac{\sigma_o^2}{\sigma^2 + \sigma_o^2}(\mu - x), \qquad (2)$$

with MMSE $L(\hat{\theta}_G; G) = \frac{\sigma^2 \sigma_o^2}{\sigma^2 + \sigma_o^2} = \sigma_o^2 - \frac{\sigma_o^4}{\sigma^2 + \sigma_o^2}$.

First, we consider the simplest case of Gaussian population distribution $\mathcal{N}(\mu, \sigma^2)$ with known variance σ^2 but unknown mean μ. The following theorem characterizes the minimax regret of this case. This case has been previously studied in [5]. We use a general hierarchical-Bayes-with-hyper-prior approach, which motivates the estimators for the vector case.

Theorem III.1. *Given a known variance σ^2 and the family of population distributions $\mathcal{G} = \{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\}$, the minimax regret w.r.t. to the family $\mathcal{G}$ is*

$$\text{TotReg}(\mathcal{G}) = \frac{\sigma_o^4}{\sigma^2 + \sigma_o^2}.$$

Proof of Theorem III.1. The upper bound is achieved by estimator

$$\hat{\theta}_i(x_{1:m}) = x_i + \frac{\sigma_o^2}{\sigma^2 + \sigma_o^2}(\bar{x}_{1:m} - x_i), \qquad (3)$$

where $\bar{x}_{1:m} = \frac{\sum_{i=1}^m x_i}{m}$. We have $\text{TotReg}(\hat{\theta}_{1:m}; \mathcal{N}(\mu, \sigma^2)) = \frac{\sigma_o^4}{\sigma^2 + \sigma_o^2}$, and thus $\text{TotReg}(\mathcal{G}) \leq \frac{\sigma_o^4}{\sigma^2 + \sigma_o^2}$.

The exactly tight lower bound is proved by considering a hierarchical Bayes model with hyper prior $\mu \sim \mathcal{N}(\nu, s^2)$ for some ν, s^2. We can calculate that $\theta_i | \mathrm{x}_{1:m} \sim \mathcal{N}(\hat{\theta}_i^H, (\hat{\sigma}^H)^2)$,

where the posterior mean $\hat{\theta}_i^H$ is the Bayesian optimal estimator of θ_i with closed-form

$$\hat{\theta}_i^H = \mathrm{x}_i + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\left(\bar{\mathbf{x}}_{1:m} - \mathrm{x}_i\right) + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\frac{\frac{\sigma^2+\sigma_o^2}{m}}{s^2+\frac{\sigma^2+\sigma_o^2}{m}}\left(\nu - \bar{\mathbf{x}}_{1:m}\right). \tag{4}$$

and the posterior variance $(\hat{\sigma}_i^H)^2$ is

$$(\hat{\sigma}_i^H)^2 = \frac{\sigma^2\sigma_o^2}{\sigma^2+\sigma_o^2} + \frac{\sigma_o^4}{\sigma^2+\sigma_o^2}\frac{s^2}{ms^2+\sigma^2+\sigma_o^2}.$$

It is straightforward to see that $(\hat{\sigma}_i^H)^2$ is the mean square error and actually the minimum mean square error given hyper prior $\mathcal{N}(\nu, s^2)$. We know $\mathrm{TotReg}(\mathcal{G}) \geq \mathbb{E}[L(\theta_{1:m}^H; G) - mL(\hat{\theta}_G; G)] \geq \sum_{i=1}^m (\hat{\sigma}_i^H)^2 - m\frac{\sigma^2\sigma_o^2}{\sigma^2+\sigma_o^2}$ for any arbitrary ν, s^2, where the expectation is taken w.r.t. $\mu \sim \mathcal{N}(\nu, s^2)$. We then conclude the lower bound that

$$\mathrm{TotReg}(\mathcal{G}) \geq \sup_{\nu, s^2} \frac{\sigma_o^4}{\sigma^2+\sigma_o^2}\frac{ms^2}{ms^2+\sigma^2+\sigma_o^2} = \frac{\sigma_o^4}{\sigma^2+\sigma_o^2}, \tag{5}$$

where the supremum is taken by any finite ν and $s^2 \to \infty$. □

Remark: A lower bound of $\mathrm{TotReg}(\mathcal{G}) \geq \frac{\sigma_o^4}{8(\sigma^2+\sigma_o^2)}$ can be proved via Le Cam's two-point method in Appendix A, while such a method, though can give an order-wise optimal bound, does not characterize the exact multiplicative constant.

According to Theorem III.1, larger σ^2 leads to smaller regret. It is slightly counter-intuitive since the overall error is increasing as σ^2 increases. But it is also easy to interpret, since the impact of others' observations on the local estimation is based on the prior and is getting weaker with larger σ^2, thus the estimator is more competitive with the MMSE.

We then consider the general case in which the variance σ^2 is also unknown to the estimator. James-Stein estimator [2] is $\hat{\theta}_{1:m}^{\mathrm{JS}}(\mathrm{x}_{1:m})$ with

$$\hat{\theta}_i^{\mathrm{JS}}(\mathrm{x}_{1:m}) = \mathrm{x}_i + \frac{\sigma_o^2(m-3)}{\|\mathrm{x}_{1:m} - \mathbf{1}_m\bar{\mathrm{x}}_{1:m}\|^2}(\bar{\mathrm{x}}_{1:m} - \mathrm{x}_i), \tag{6}$$

where $\mathbf{1}_m$ is a m-length vector with all ones. It resembles (3) by estimating a contraction factor $\frac{\sigma_o^2}{\sigma^2+\sigma_o^2}$ by $\frac{\sigma_o^2(m-3)}{\|\mathrm{x}_{1:m}-\mathbf{1}\bar{\mathrm{x}}\|^2}$.

Theorem III.2. *James-Stein estimator, without the knowledge of σ^2, is minimax order-optimal with regret*

$$\mathrm{TotReg}(\hat{\theta}_{1:m}^{JS}; \mathcal{G}) = \frac{3\sigma_o^4}{\sigma^2+\sigma_o^2}.$$

Proof of Theorem III.2. The MSE of James-Stein estimator is $L(\hat{\theta}_{1:m}^{\mathrm{JS}}; G) = \frac{m\sigma_o^2\sigma^2}{\sigma_o^2+\sigma^2} + \frac{3\sigma_o^4}{\sigma_o^2+\sigma^2}$. It is calculated by Stein's unbiased risk estimator (SURE) and the detailed computation is given in the Appendix C. Since the MSE of the MMSE estimator (2) is $L(\hat{\theta}_G; G) = \frac{\sigma_o^2\sigma^2}{\sigma_o^2+\sigma^2}$, the total regret is thus $\frac{3\sigma_o^4}{\sigma_o^2+\sigma^2}$. The upper bound matches the lower bound for the known variance case in Eq (5) within a constant factor of 3 and thus minimax order-optimal. Moreover, JS estimator is actually minimax (in terms of μ) order-wise optimal in a universal manner (in terms of σ^2). □

B. Vector case

The unknown parameter $\boldsymbol{\theta}_i \in \mathbb{R}^d$ with a Gaussian population distribution $G = \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d)$. We have observations $\mathbf{x}_i|\boldsymbol{\theta}_{1:m} \sim \mathcal{N}(\boldsymbol{\theta}_i, \sigma_o^2\mathbb{I}_d)$ for each $i = 1, \ldots, m$. Similar to the scalar case, the MMSE estimator given G is $\hat{\boldsymbol{\theta}}_G(\mathbf{x}) = \mathbf{x} + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}(\boldsymbol{\mu} - \mathbf{x})$, with MSE $L(\hat{\boldsymbol{\theta}}_G; G) = d\frac{\sigma^2\sigma_o^2}{\sigma^2+\sigma_o^2}$.

We start from the simple case of multivariate Gaussian population distribution $\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d)$ with known σ^2 and characterize the minimax regret of this case in the theorem below.

Theorem III.3. *Given a known variance σ^2 and the family of population distributions $\mathcal{G} = \{\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d) : \boldsymbol{\mu} \in \mathbb{R}^d\}$, the minimax regret w.r.t. the family $\mathcal{G}$ is*

$$\mathrm{TotReg}(\mathcal{G}) = d\frac{\sigma_o^4}{\sigma^2+\sigma_o^2}.$$

Proof of Theorem III.3. The lower bound of the worst-case total regret $\mathrm{TotReg}(\mathcal{G}) \geq d\frac{\sigma_o^4}{\sigma^2+\sigma_o^2}$ can be obtained applying (5) at each coordinate. A naive estimator is applying the estimator Eq (3) at each coordinate and it is indeed minimax optimal. We here propose a hierarchical James-Stein estimator with known variance σ^2, which is more adaptive and dominates this naive estimator by aggregating the information across coordinates in James-Stein's manner, as follows. For user-i at coordinate k,

$$\hat{\boldsymbol{\theta}}_{i,k}^{\mathrm{HJS},\sigma^2} = \mathbf{x}_{i,k} + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\left(\bar{\mathbf{x}}_k - \mathbf{x}_{i,k}\right) + \frac{\sigma_o^2}{m}\frac{d-3}{\|\bar{\mathbf{x}}_{1:d} - \bar{\mathbf{x}}\mathbf{1}_d\|^2}\left(\bar{\mathbf{x}} - \bar{\mathbf{x}}_k\right), \tag{7}$$

where $\bar{\mathbf{x}}_k = \frac{\sum_{i=1}^m \mathbf{x}_{i,k}}{m}$ is the sample average of coordinate k and $\bar{\mathbf{x}} = \frac{\sum_{k=1}^d \bar{\mathbf{x}}_k}{d} = \frac{\sum_{i,k}\mathbf{x}_{i,k}}{md}$ is the overall average. The estimator $\hat{\boldsymbol{\theta}}_{i,k}^{\mathrm{HJS},\sigma^2}$ is motivated by the hierarchical Bayes estimator in Eq (4), as they share the same first and second terms and the last term is a resemble of the last term $\frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\frac{\frac{\sigma^2+\sigma_o^2}{m}}{s^2+\frac{\sigma^2+\sigma_o^2}{m}}(\nu - \bar{\mathbf{x}}_{1:m}) = \frac{\sigma_o^2}{m}\frac{m}{ms^2+\sigma^2+\sigma_o^2}(\nu - \bar{\mathbf{x}}_{1:m})$ of Eq (4) by estimating $\frac{m}{ms^2+\sigma^2+\sigma_o^2}$ by $\frac{d-3}{\|\bar{\mathbf{x}}_{1:d}-\bar{\mathbf{x}}\mathbf{1}_d\|^2}$ in *James-Stein manner across coordinates* and ν by $\bar{\mathbf{x}}$. The MSE of this estimator can be calculated as

$$L(\hat{\boldsymbol{\theta}}_{1:m}^{\mathrm{HJS},\sigma^2}; \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d)) = md\frac{\sigma^2\sigma_o^2}{\sigma^2+\sigma_o^2} + d\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} - \frac{d-3}{m}\mathbb{E}\left[\frac{(d-3)\sigma_o^4}{\|\bar{\mathbf{x}}_{1:d} - \bar{\mathbf{x}}\mathbf{1}_d\|^2}\right],$$

and thus its regret is $\mathrm{TotReg}(\hat{\boldsymbol{\theta}}_{1:m}^{\mathrm{HJS},\sigma^2}; \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d)) = d\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} - \frac{d-3}{m}\mathbb{E}\left[\frac{(d-3)\sigma_o^4}{\|\bar{\mathbf{x}}_{1:d}-\bar{\mathbf{x}}\mathbf{1}_d\|^2}\right]$.

Further suppose that the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ satisfies $\mu_1, \ldots, \mu_d \overset{i.i.d.}{\sim} \mathcal{N}(\nu, s^2)$ as in the hierarchical Bayes model, we then have $\mathbb{E}_{\nu,s^2}\left[\frac{(d-3)\sigma_o^4}{\|\bar{\mathbf{x}}_{1:d}-\bar{\mathbf{x}}\mathbf{1}_d\|^2}\right] = m\frac{\sigma_o^4}{ms^2+\sigma^2+\sigma_o^2}$, which implies

$$\mathbb{E}_{\nu,s^2}\left[L(\hat{\boldsymbol{\theta}}_{1:m}^{\mathrm{HJS},\sigma^2}; \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d))\right] = md\frac{\sigma^2\sigma_o^2}{\sigma^2+\sigma_o^2} + d\frac{\sigma_o^4}{\sigma^2+\sigma_o^2}\left(1 - \frac{\sigma^2+\sigma_o^2}{ms^2+\sigma^2+\sigma_o^2}\right) + 3\frac{\sigma_o^4}{ms^2+\sigma^2+\sigma_o^2}.$$

When $s = 0$, i.e., $\mu_1 = \cdots = \mu_d$, we know

$$\mathbb{E}_{\nu,s^2}\left[L(\hat{\boldsymbol{\theta}}^{\text{HJS},\sigma^2}_{1:m}; \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d))\right] = \frac{md\sigma^2\sigma_o^2}{\sigma^2+\sigma_o^2} + \frac{3\sigma_o^4}{\sigma^2+\sigma_o^2}.$$

When $s \to \infty$, we know

$$\mathbb{E}_{\nu,s^2}[L(\hat{\boldsymbol{\theta}}^{\text{HJS},\sigma^2}_{1:m}; \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d))] = \frac{md\sigma^2\sigma_o^2}{\sigma^2+\sigma_o^2} + \frac{d\sigma_o^4}{\sigma^2+\sigma_o^2},$$

which achieves the *worst-case* lower bound suggested by applying hierarchical Bayes at each coordinate. □

Remark: Note that the proposed $\hat{\boldsymbol{\theta}}^{\text{HJS},\sigma^2}$ is more adaptive than using estimator Eq (3) for each coordinates. Though they both achieve the minimax optimal regret in worst-case, when the unknown variance of the hyper prior s is small (coordinates $\mu_1, \ldots, \mu_d$ are close), $\hat{\boldsymbol{\theta}}^{\text{HJS},\sigma^2}$ can have smaller dimension-independent regret while the latter fails to take advantage of the structure among coordinates with regret linear in d.

We then consider the case where σ^2 is unknown to the estimator. Motivated by the hierarchical James-Stein estimator with known variance in Eq (7), we propose the hierarchical James-Stein estimator below

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{\text{HJS}}_{i,k} = \mathbf{x}_{i,k} &+ \frac{\sigma_o^2(d(m-1)-2)}{\sum_{h=1}^d \|\mathbf{x}_{1:m,h} - \bar{\mathbf{x}}_h \mathbf{1}_m\|^2}(\bar{\mathbf{x}}_k - \mathbf{x}_{i,k}) \\ &+ \frac{\sigma_o^2}{m}\frac{d-3}{\|\bar{\mathbf{x}}_{1:d} - \bar{\mathbf{x}}\mathbf{1}_d\|^2}(\bar{\mathbf{x}} - \bar{\mathbf{x}}_k),\end{aligned} \tag{8}$$

where the second term is an estimate of $\frac{\sigma_o^2}{\sigma^2+\sigma_o^2}(\bar{\mathbf{x}}_k - \mathbf{x}_{i,k})$ in James-Stein manner.

Theorem III.4. *Hierarchical James-Stein estimator, without the knowledge of σ^2, is minimax order-optimal with regret*

$$\begin{aligned}&\textit{TotReg}(\hat{\boldsymbol{\theta}}^{\textit{HJS}}_{1:m}; \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d)) \\ &= d\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} + 2\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} - \frac{d-3}{m}\mathbb{E}\left[\frac{(d-3)\sigma_o^4}{\|\bar{\mathbf{x}}_{1:d} - \bar{\mathbf{x}}\mathbf{1}_d\|^2}\right].\end{aligned}$$

The detailed calculation is in the Appendix E. Since the last term of the regret is negative, the extra regret due to not accessing σ^2 is within $2\frac{\sigma_o^4}{\sigma^2+\sigma_o^2}$, which is dimension independent, and thus it is minimax order-optimal. Moreover, suppose $\mu_1, \ldots, \mu_d \sim \mathcal{N}(\nu, s^2)$, we have

$$\begin{aligned}&\mathbb{E}_{\nu,s^2}\left[\text{TotReg}(\hat{\boldsymbol{\theta}}^{\text{HJS}}_{1:m}; \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d))\right] \\ &= d\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} + 2\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} - (d-3)\frac{\sigma_o^4}{ms^2+\sigma^2+\sigma_o^2}.\end{aligned}$$

When $s \to 0$, even though the estimator does not have access to s, the expected regret is dimension-independent $5\frac{\sigma_o^4}{\sigma^2+\sigma_o^2}$. In contrast, the estimator without taking advantage of collaboration among coordinates will scale linearly w.r.t. d.

IV. Personalized estimation with communication constraints

We consider the vector case with the unknown parameter $\boldsymbol{\theta}_i \in \mathbb{R}^d$ following population distribution $G = \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d)$, and each user-$i$ has n local observations $\mathbf{x}_i^{(1:n)}|\boldsymbol{\theta}_{1:m} \overset{i.i.d.}{\sim} \mathcal{N}(\boldsymbol{\theta}_i, \sigma_o^2\mathbb{I}_d)$, $i = 1, \ldots, m$. Moreover, we have communication constraints that the codeword $Z_i = f_i(\mathbf{x}_i^{(1:n)}, U_i) \in \mathcal{Z} \subseteq \{0,1\}^*$ from any user-i has expected length at most B.

Let $\mathbf{x}_i = \frac{\sum_{j=1}^n \mathbf{x}_i^{(j)}}{n}$ be the average of data $\mathbf{x}_i^{(1:n)}$. Since $\mathbf{x}_i$ is a sufficient statistic for $\boldsymbol{\theta}_i$ and Z_i is a function of $\mathbf{x}_i^{(1:n)}$, we have the Markov chain $\boldsymbol{\theta}_i - \mathbf{x}_i - \mathbf{x}_i^{(1:n)} - Z_i$. Thus without loss of generality, we simply consider $\mathbf{x}_i$ as user-i's data with $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\theta}_i, \sigma_o^2/n)$ and $Z_i = f_i(\mathbf{x}_i, U_i)$.

This following theorem shows that it is both sufficient and necessary to have communication $B = \tilde{\Theta}(d)$ to achieve the best total regret as in an unconstrained setting, and thus with B bits of message from each user, the collaboration among users improves personalized estimation.

Theorem IV.1. *Given a family of population distributions $\mathcal{G} = \{\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d) : \|\boldsymbol{\mu}\|^2 \le d\}$ with some bounded σ^2, the minimax regret w.r.t. the family $\mathcal{G}$ for some $B = \tilde{\Theta}(d)$ is*

$$\text{TotReg}(\mathcal{G}, B) = \Theta\left(d\frac{(\sigma_o^2/n)^2}{\sigma_o^2/n + \sigma^2}\right).$$

Remark: The achievability (Theorem IV.3) takes $B = \tilde{\Omega}(d)$, and the lower bound (Theorem IV.4) explicitly depends on B with a wider range. The characterization of the regret for the entire range of (n, σ^2, d, B) is left for future work.

A. Estimation via Poisson representation

We use the Poisson functional representation proposed by Li and Gamal [4]. User-i sends a message of expected finite length to a receiver (i.e., other users) and the receiver can "simulate" a noisy version of the the user's local observation $\mathbf{x}_i$. We take one user-i as an example and omit the index i for simplicity. Specifically, the external noise U, encoding function f at the sender side, and the decoding at the receiver side are illustrated as follows with two coefficients $\sigma_q^2, \lambda > 1$ specified later.

- External randomness $U = \{(Y_k, T_k)\}_{k\ge1}$ is a marked Poisson point process (PPP), where $0 \le T_1 \le T_2 \le \cdots$ is a PPP with rate 1 and $Y_1, Y_2, \ldots \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_q^2\mathbb{I}_d)$. It is a common randomness generated on both the sender and receiver sides.
- Encoding $Z = f(\mathbf{x}, U)$, where $\mathbf{x}$ is the sender's local observation, is by two steps.
  1) Compute an index $K = \arg\min_k T_k \cdot \frac{\mathcal{N}(\mathbf{y}_k; 0, \sigma_q^2\mathbb{I}_d)}{\mathcal{N}(\mathbf{y}_k; \mathbf{x}, \sigma_s^2\mathbb{I}_d)}$.
  2) Encode K into Z by an optimal prefix-free code for the Zipf distribution $q(k) \propto k^{-\lambda}$ for $k \ge 1$.
- Decoding $\mathbf{y} = g(Z, U) = Y_K$ with K decoded from Z.

Lemma IV.2 (Channel simulation [4])**.** *Transmitting $\mathbf{x}$ according to the above procedure with any σ_q^2 and $\lambda > 1$, the recovered signal $\mathbf{y}$ follows distribution $\mathcal{N}(\mathbf{x}, \sigma_s^2\mathbb{I}_d)$.*

The lemma above characterizes the quality of the decoded signal $\mathbf{y}_i$ if user-i holds local observation $\mathbf{x}_i$. As discussed in [4], the above procedure can be viewed as channel simulation, i.e., the receiver simulates the output $\mathbf{y}$ of a Gaussian channel with input $\mathbf{x}$.

We present the total MSE of different estimators as follows. The first two are MMSE estimators with population distribution $G = \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \mathbb{I}_d)$ or hyper prior $\boldsymbol{\mu} \overset{i.i.d.}{\sim} \mathcal{N}(\nu, s^2)$. The middle three MSE's are taken expectation w.r.t. to the hyper prior, and they decrease as s^2 decreases. The maximum likelihood estimator (MLE) at the bottom is $\hat{\boldsymbol{\theta}}_i = \frac{\sum_{j=1}^n \mathbf{x}_i^{(j)}}{n}$ without collaboration. From top to bottom, the MSE is increasing, and thus collaboration improves personalized estimation.

$$
\begin{aligned}
L(\hat{\boldsymbol{\theta}}_{1:m}^{\mathrm{MMSE},G}; G) &= \frac{md\sigma_o^2}{n} - \frac{md\sigma_o^4}{n^2\sigma^2 + n\sigma_o^2} \\
\mathbb{E}_{\nu,s^2}[L(\hat{\boldsymbol{\theta}}_{1:m}^{\mathrm{MMSE},\mathcal{N}(\nu,s^2)}; G)] &= \frac{md\sigma_o^2}{n} - \frac{md\sigma_o^4}{n^2\sigma^2 + n\sigma_o^2} + \frac{md\sigma_o^4 s^2}{(n\sigma^2 + \sigma_o^2)(mns^2 + n\sigma^2 + \sigma_o^2)} \\
\mathbb{E}_{\nu,s^2}[L(\hat{\boldsymbol{\theta}}_{1:m}^{\mathrm{HJS},\sigma^2}; G)] &= \frac{md\sigma_o^2}{n} - \frac{md\sigma_o^4}{n^2\sigma^2 + n\sigma_o^2} + \frac{md\sigma_o^4 s^2}{(n\sigma^2 + \sigma_o^2)(mns^2 + n\sigma^2 + \sigma_o^2)} + \frac{3\sigma_o^4/n}{(ms^2+\sigma^2)n + \sigma_o^2} \\
\mathbb{E}_{\nu,s^2}[L(\hat{\boldsymbol{\theta}}_{1:m}^{\mathrm{HJS}}; G)] &= \frac{md\sigma_o^2}{n} - \frac{md\sigma_o^4}{n^2\sigma^2 + n\sigma_o^2} + \frac{md\sigma_o^4 s^2}{(n\sigma^2 + \sigma_o^2)(mns^2 + n\sigma^2 + \sigma_o^2)} + \frac{3\sigma_o^4/n}{(ms^2+\sigma^2)n + \sigma_o^2} + \frac{2\sigma_o^4}{n^2\sigma^2 + n\sigma_o^2} \\
L(\hat{\boldsymbol{\theta}}_{1:m}^{\mathrm{MLE}}; G) &= \frac{md\sigma_o^2}{n}
\end{aligned}
$$

Theorem IV.3. *For the family of population distribution $\mathcal{G} = \{\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d) : \|\boldsymbol{\mu}\|_2 \le \sqrt{d}D\}$ with $\sigma \le S$, and under communication budget $B = \tilde{\Omega}(b)$ where $b = \frac{d}{2}\log\left(1 + \frac{(\sigma_o^2/n + D^2 + S^2)S^2}{\sigma_o^4} n^2 m\right)$, the minimax regret w.r.t. $\mathcal{G}$ is upper bounded by*

$$
\mathrm{TotReg}(\mathcal{G}, B) = O\left(d\frac{(\sigma_o^2/n)^2}{\sigma_o^2/n + \sigma^2}\right).
$$

Proof of Theorem IV.3. Let $\sigma_n^2 := \sigma_o^2/n$. Take $\sigma_s^2 = \frac{\sigma_n^4}{S^2 m}$, $\sigma_q^2 = \sigma_s^2 + \sigma_n^2 + D^2 + S^2$ and $\lambda = 1 + 1/(b + e^{-1}\log e + 1)$ for the Poisson functional representation communication procedure. Transmitting each user's local observations $\{\mathbf{x}_i : i = 1, \ldots, m\}$ to other users, and each user can recover signals $\{\mathbf{y}_i : i = 1, \ldots, m\}$. Note that $\mathbf{y}_i \sim \mathcal{N}(\boldsymbol{\theta}_i, (\sigma_n^2 + \sigma_s^2)\mathbb{I}_d)$ by Lemma IV.2, the difference between the minimum MSEs w.r.t. data $\mathbf{y}_{1:m}$ and $\mathbf{x}_{1:m}$ normalized by d is $\frac{\sigma^2(\sigma_n^2+\sigma_s^2)}{\sigma^2+\sigma_n^2+\sigma_s^2} - \frac{\sigma^2\sigma_n^2}{\sigma^2+\sigma_n^2} = \frac{\sigma_s^2}{\sigma^2+\sigma_n^2}\frac{\sigma^4}{\sigma^2+\sigma_n^2+\sigma_s^2} \le \frac{S^2\sigma_s^2}{\sigma^2+\sigma_n^2} \le \frac{\sigma_n^4/m}{\sigma^2+\sigma_n^2}$. We can apply $\hat{\theta}^{\mathrm{HSJ}}$ Eq (8) by data $\{\mathbf{y}_i\}$, and the corresponding regret upper bound w.r.t. data $\mathbf{y}_{1:m}$ can be calculated as in Theorem III.4 by replacing the observation noise from σ_o^2 to $\sigma_n^2 + \sigma_s^2 = O(\sigma_n^2)$. The regret is thus upper bounded by $O\left(d\frac{(\sigma_o^2/n)^2}{\sigma_o^2/n+\sigma^2}\right)$.

Let $\mathrm{Q}(\cdot) = \mathcal{N}(\cdot; 0, \sigma_q^2\mathbb{I}_d)$ and $\mathrm{P}_{\mathbf{y}|\mathbf{x}}(\cdot|\mathbf{x}) = \mathcal{N}(\cdot; \mathbf{x}, \sigma_s^2\mathbb{I}_d)$. It has been shown in [4, Sketch of the proof of Theorem 1] that the computed index $K = \arg\min_k T_k \cdot \frac{\mathrm{P}_{\mathbf{y}|\mathbf{x}}(Y_k|\mathbf{x})}{\mathrm{Q}(Y_k)}$ satisfies $\mathbb{E}[\log K] \le \mathbb{E}[\mathrm{KL}(\mathrm{P}_{\mathbf{y}|\mathbf{x}}(\cdot|\mathbf{x})||\mathrm{Q})] + e^{-1}\log e + 1$. Note that

$$
\mathrm{KL}(\mathrm{P}_{\mathbf{y}|\mathbf{x}}(\cdot|\mathbf{x})||\mathrm{Q}) = \left(\frac{\|\mathbf{x}\|^2}{2\sigma_q^2} + \frac{d\sigma_s^2}{2\sigma_q^2} - \frac{d}{2}\right)\log e + \frac{d}{2}\log\frac{\sigma_q^2}{\sigma_s^2}.
$$

Since $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, (\sigma^2+\sigma_n^2+\sigma_s^2)\mathbb{I}_d)$ and $\sigma_q^2 = \sigma_s^2+\sigma_n^2+D^2+S^2$, we have

$$
\begin{aligned}
&\mathbb{E}[\mathrm{KL}(\mathrm{P}_{\mathbf{y}|\mathbf{x}}(\cdot|\mathbf{x})||\mathrm{Q})] \\
&= \left(\frac{\|\boldsymbol{\mu}\|^2 + d\sigma^2 + d\sigma_n^2 + d\sigma_s^2}{2\sigma_q^2} - \frac{d}{2}\right)\log e + \frac{d}{2}\log\frac{\sigma_q^2}{\sigma_s^2} \\
&\le \frac{d}{2}\log\frac{\sigma_q^2}{\sigma_s^2} = \frac{d}{2}\log\left(1 + \frac{(\sigma_n^2 + D^2 + S^2)S^2}{\sigma_o^4} n^2 m\right) = b,
\end{aligned}
$$

where the inequality is by $\|\boldsymbol{\mu}\| \le D, \sigma \le S$ and $\frac{\|\boldsymbol{\mu}\|^2 + d\sigma^2 + d\sigma_n^2 + d\sigma_s^2}{2\sigma_q^2} \le \frac{d}{2}$. Then as shown in [4], encoding K into Z by an optimal prefix-free code for the Zipf distribution $q(k) \propto k^{-\lambda}$ with $\lambda = 1 + 1/(b + e^{-1}\log e + 1)$ satisfies $\mathbb{E}[|Z|] \le b + \log(b+1) + 5$. We thus have $B \ge b + \log(b+1) + 5$, i.e., $B = \tilde{\Omega}(b)$. □

B. Lower bound of regret under communication constraints

The lower bound analysis of the regret follows a similar logic as that in [8] with important definitions such as strong data processing inequalities. Zhang et. al. [8] considered the setting where each user holds i.i.d. samples from a common distribution parameterized by $\boldsymbol{\theta}$, and the goal is to estimate this parameter $\boldsymbol{\theta}$. In our work, the data across users are heterogeneous and the goal is personalized estimation where each user-i is estimating a local parameter $\boldsymbol{\theta}_i$. The proof of Theorem IV.4 is given in Appendix F.

Theorem IV.4. *Given a known variance σ^2 and the family of population distributions $\mathcal{G} = \{\mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbb{I}_d) : \|\boldsymbol{\mu}\|^2 \le d\}$, the minimax regret $\mathrm{TotReg}(\mathcal{G}, B)$ under communication budget B is lower bounded by*

$$
\Omega\left(d\frac{(\frac{\sigma_o^2}{n})^2}{\frac{\sigma_o^2}{n} + \sigma^2}\min\left(\frac{m}{\frac{\sigma_o^2}{n} + \sigma^2}, \frac{m}{\log m}, \frac{d}{\min(B\log m, d)}\right)\right).
$$

V. Conclusion

We study the personalized Gaussian mean estimation under communication constraints. The worst-case regret is characterized exactly without communication constraints when the variance of the population distribution is known and order-wise optimal regret upper bounds up to multiplicative constant 3 are provided. We show that when the message sent by each user has $B = \tilde{\Theta}(d)$ bits, the order-wise optimal worst-case regret $O\left(d\frac{(\sigma_o^2/n)^2}{\sigma_o^2/n+\sigma^2}\right)$ is the same as that without communication constraints.

ACKNOWLEDGMENT

The work was supported in part by NSF grants 2139304, 2007714, 2146838, and the Army Research Laboratory grant under Cooperative Agreement W911NF-17-2-0196.

REFERENCES

[1] J. C. Duchi and M. J. Wainwright. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

[2] W. James and C. Stein. Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. on Math. Statist. and Prob., 1961*, 1961.

[3] W. Jiang and C.-H. Zhang. General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics*, 37(4):1647–1684, 2009.

[4] C. T. Li and A. El Gamal. Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978, 2018.

[5] K. Ozkara, A. M. Girgis, D. Data, and S. Diggavi. A statistical framework for personalized federated learning and estimation: Theory, algorithms, and privacy. In *The Eleventh International Conference on Learning Representations*, 2023.

[6] Y. Polyanskiy and Y. Wu. Sharp regret bounds for empirical bayes and compound decision problems. *arXiv preprint arXiv:2109.03943*, 2021.

[7] B. Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

[8] Y. Zhang, J. Duchi, M. I. Jordan, and M. J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. *Advances in Neural Information Processing Systems*, 26, 2013.

APPENDIX

DETAILED CALCULATION FOR THE SCALAR CASE

A. Lower bound via Le Cam's method for Thm III.1

As discussed in the Proof of Theorem III.1, we can also use the Le Cam's two-point method [7] to prove the lower bound. Given the separation condition w.r.t. $G = \mathcal{N}(\mu, \sigma^2)$ and $G' = \mathcal{N}(\mu', \sigma^2)$,

$$\begin{aligned}
&\inf_{a_{1:m}} \left\{ \|a_{1:m} - \hat{\theta}_G(\mathrm{x}_{1:m})\|^2 + \|a_{1:m} - \hat{\theta}_{G'}(\mathrm{x}_{1:m})\|^2 \right\} \\
&\geq \frac{1}{2} \|\hat{\theta}_G(\mathrm{x}_{1:m}) - \hat{\theta}_{G'}(\mathrm{x}_{1:m})\|^2 \\
&= \frac{m\sigma_o^4(\mu-\mu')^2}{2(\sigma^2+\sigma_o^2)^2} =: \Delta.
\end{aligned}$$

We then have

$$\begin{aligned}
&\mathrm{TotReg}(\hat{\theta}_{1:m}; \mathcal{G}) \geq \max_{G \in \{\mathcal{N}(\mu,\sigma^2), \mathcal{N}(\mu',\sigma^2)\}} \mathrm{TotReg}(\hat{\theta}_{1:m}; G) \\
&\geq \frac{\Delta}{2} \left(1 - \|P_{\mathrm{x}_{1:m}}(\cdot; G) - P_{\mathrm{x}_{1:m}}(\cdot; G')\|_{\mathrm{TV}}\right) \\
&\geq \frac{m\sigma_o^4(\mu-\mu')^2}{4(\sigma^2+\sigma_o^2)^2} \Bigg(1 - \\
&\quad \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathcal{N}(\mu, \sigma^2+\sigma_o^2)^{\otimes m}; \mathcal{N}(\mu', \sigma^2+\sigma_o^2)^{\otimes m}\right)}\Bigg) \\
&= \frac{m\sigma_o^4(\mu-\mu')^2}{4(\sigma^2+\sigma_o^2)^2} \left(1 - \sqrt{\frac{m(\mu-\mu')^2}{4(\sigma^2+\sigma_o^2)}}\right).
\end{aligned}$$

Taking $\mu' = \mu + \sqrt{\frac{\sigma^2+\sigma_o^2}{m}}$ concludes the proof that $\mathrm{TotReg}(\hat{\theta}_{1:m}; \mathcal{G}) \geq \frac{\sigma_o^4}{8(\sigma^2+\sigma_o^2)}$, which is order optimal.

Note if we have more local samples from each user, e.g., $P_{\mathrm{x}_i^{(1:n)}}(\cdot; G) = \mathcal{N}(\mu, \sigma_o^2 \mathbb{I}_n + \sigma^2 \mathbf{1}\mathbf{1}^\top)$. We can still calculate KL-divergence $\mathrm{KL}\left(P_{\mathrm{x}_i^{(1:n)}}(\cdot; G); P_{\mathrm{x}_i^{(1:n)}}(\cdot; G')\right) = \frac{(\mu-\mu')^2}{2(\sigma^2+\sigma_o^2/n)}$ and then derive a lower bound of $\frac{(\sigma_o^2/n)^2}{8(\sigma^2+\sigma_o^2/n)}$.

B. Calculation for the hierarchical Bayesian model in (4)

We consider a hierarchical Bayesian model with hyper prior $\mu \sim \mathcal{N}(\nu, s^2)$ for some ν, s^2. We can calculate that $\mu|\mathrm{x}_{2:n} \sim \mathcal{N}\left(\frac{s^2}{\frac{\sigma^2+\sigma_o^2}{m-1}+s^2}\frac{\sum_{i=2}^m \mathrm{x}_i}{m-1} + \frac{\frac{\sigma^2+\sigma_o^2}{m-1}}{\frac{\sigma^2+\sigma_o^2}{m-1}+s^2}\nu, \alpha_s\right)$, where $\alpha_s = \left(\frac{m-1}{\sigma^2+\sigma_o^2} + \frac{1}{s^2}\right)^{-1}$. Thus $\theta_1|\mathrm{x}_{2:n} = x_{2:n} \sim \mathcal{N}\left(\frac{s^2}{\frac{\sigma^2+\sigma_o^2}{m-1}+s^2}\frac{\sum_{i=2}^m x_i}{m-1} + \frac{\frac{\sigma^2+\sigma_o^2}{m-1}}{\frac{\sigma^2+\sigma_o^2}{m-1}+s^2}\nu, \alpha_s + \sigma^2\right)$, and the posterior mean $\mathbb{E}[\theta_1|\mathrm{x}_{1:m}]$ can be calculated as in Eq (9).

Denote by $\hat{\theta}^H_{1:m}(\mathrm{x}_{1:m}) = \mathbb{E}[\theta_{1:m}|\mathrm{x}_{1:m}]$, which is the optimal estimator given access to the hyper prior $\mu \sim \mathcal{N}(\nu, s^2)$. Taking $s \to \infty$ while ν fixed gives

$$\begin{aligned}
\hat{\theta}^H_1(\mathrm{x}_{1:m}) &= \mathrm{x}_1 + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\left(\frac{\sum_{i=1}^m \mathrm{x}_i}{m} - \mathrm{x}_1\right) \\
&= \frac{\sigma^2}{\sigma^2+\sigma_o^2}\mathrm{x}_1 + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\bar{\mathrm{x}}_{1:m},
\end{aligned}$$

which is valid estimator without relying on any unknown parameters and coincides with the estimator in the upper bound. Taking $s = 0$ gives

$$\hat{\theta}^H_1(\mathrm{x}_{1:m}) = \mathrm{x}_1 + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}(\nu - \mathrm{x}_1), \tag{10}$$

which is $\hat{\theta}_G$ with $G = \mathcal{N}(\nu, \sigma^2)$.

Moreover, the posterior variance can be calculated as

$$\begin{aligned}
&\mathrm{Var}(\theta_1|\mathrm{x}_{1:m}) = ((\alpha_s+\sigma^2)^{-1} + \sigma_o^{-2})^{-1} \\
&= \frac{(\alpha_s+\sigma^2)\sigma_o^2}{\alpha_s+\sigma^2+\sigma_o^2} = \sigma_o^2\left(1 - \frac{\sigma_o^2/\alpha_s}{1+\frac{\sigma^2+\sigma_o^2}{\alpha_s}}\right) \\
&= \sigma_o^2\left(1 - \frac{\sigma_o^2\left(\frac{m-1}{\sigma^2+\sigma_o^2} + \frac{1}{s^2}\right)}{m + \frac{\sigma^2+\sigma_o^2}{s^2}}\right) \\
&= \sigma_o^2 - \frac{\sigma_o^4}{\sigma^2+\sigma_o^2}\frac{m-1}{m+\frac{\sigma^2+\sigma_o^2}{s^2}} - \frac{\sigma_o^4}{ms^2+\sigma^2+\sigma_o^2} \\
&= \sigma_o^2 - \frac{\sigma_o^4}{\sigma^2+\sigma_o^2}\left(1 - \frac{s^2+\sigma^2+\sigma_o^2}{ms^2+\sigma^2+\sigma_o^2}\right) - \frac{\sigma_o^4}{ms^2+\sigma^2+\sigma_o^2} \\
&= \frac{\sigma_o^2\sigma^2}{\sigma^2+\sigma_o^2} + \frac{\sigma_o^4}{\sigma^2+\sigma_o^2}\frac{s^2}{ms^2+\sigma^2+\sigma_o^2} \\
&= \sigma_o^2 - \frac{\sigma_o^4}{\sigma^2+\sigma_o^2} + \frac{1}{m}\frac{\sigma_o^4}{\sigma^2+\sigma_o^2}\left(1 - \frac{\sigma^2+\sigma_o^2}{ms^2+\sigma^2+\sigma_o^2}\right).
\end{aligned}$$

Taking $s = 0$ gives $\mathrm{Var}(\theta_1|\mathrm{x}_{1:m}) = \frac{\sigma_o^2\sigma^2}{\sigma^2+\sigma_o^2}$, which is the Bayes risk for knowing $\mu = \nu$. Taking $s = \infty$ gives $\mathrm{Var}(\theta_1|\mathrm{x}_{1:m}) = \frac{\sigma_o^2\sigma^2}{\sigma^2+\sigma_o^2} + \frac{1}{m}\frac{\sigma_o^4}{\sigma^2+\sigma_o^2}$, which is the Bayes risk of the estimator under improper prior of μ. The regret is then lower bounded by $\frac{1}{m}\frac{\sigma_o^4}{\sigma^2+\sigma_o^2}$. Thus, the estimator is indeed exactly minimax optimal.

C. Omitted calculation in the proof of Theorem III.2

The MSE of James-Stein estimator is calculated by Stein's unbiased risk estimator (SURE) as in Eq (11) - Eq (12). Since the first three equalities do not rely on the prior distribution of $\theta_{1:m}$, we have $\mathbb{E}_{\theta_{1:m}}[\|\theta_{1:m} - \hat{\theta}^{\mathrm{JS}}_{1:m}(\mathrm{x}_{1:m})\|^2] = m\sigma_o^2 - \sigma_o^4(m-3)^2\mathbb{E}_{\theta_{1:m}}\left[\frac{1}{\|\mathrm{x}_{1:m} - \mathbf{1}_m\bar{\mathrm{x}}\|^2}\right]$, where $\mathbb{E}_{\theta_{1:m}}$ indicates taking expectation w.r.t. the randomness of data following distributions $\{\mathcal{N}(\theta_i, \sigma_o^2)\}$. It is also known as (frequentist's) risk. The James-Stein estimator thus dominates the maximum likelihood estimator with risk $m\sigma_o^2$.

DETAILED CALCULATION FOR THE VECTOR CASE

D. Omitted calculation in the proof of Theorem III.3

We analyze the MSE of the Hierarchical JS estimator with known variance $\hat{\boldsymbol{\theta}}^{\mathrm{HJS},\sigma^2}_{i,k}$ by Stein's unbiased risk estimator (SURE). Since

$$\sum_{i,k}\frac{\partial}{\partial \mathrm{x}_{i,k}}\left(\frac{\sigma_o^2}{m}\frac{d-3}{\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}_d\|^2}(\bar{\mathrm{x}} - \bar{\mathrm{x}}_k)\right) = -\frac{\sigma_o^2(d-3)^2}{m\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}_d\|^2},$$

we have by SURE that

$$\mathbb{E}_G[\|\boldsymbol{\theta}_{1:m} - \hat{\boldsymbol{\theta}}^{\mathrm{HJS},\sigma^2}_{1:m}\|^2] = \mathbb{E}\left[md\sigma_o^2 + \|\hat{\boldsymbol{\theta}}^{\mathrm{HJS},\sigma^2}_{1:m,1:d} - \mathrm{x}_{1:m,1:d}\|^2\right]$$

$$
\begin{aligned}
\mathbb{E}[\theta_1|\mathrm{x}_{1:m}] &= \frac{\alpha_s+\sigma^2}{\alpha_s+\sigma^2+\sigma_o^2}\mathrm{x}_1 + \frac{\sigma_o^2}{\alpha_s+\sigma^2+\sigma_o^2}\left(\frac{s^2}{\frac{\sigma^2+\sigma_o^2}{m-1}+s^2}\frac{\sum_{i=2}^m x_i}{m-1} + \frac{\frac{\sigma^2+\sigma_o^2}{m-1}}{\frac{\sigma^2+\sigma_o^2}{m-1}+s^2}\nu\right)\\
&= \mathrm{x}_1 + \frac{\frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\frac{1}{\alpha_s}}{\frac{m}{\sigma^2+\sigma_o^2}+\frac{1}{s^2}}\left(\frac{s^2}{\frac{\sigma^2+\sigma_o^2}{m-1}+s^2}\frac{\sum_{i=2}^m x_i}{m-1} + \frac{\frac{\sigma^2+\sigma_o^2}{m-1}}{\frac{\sigma^2+\sigma_o^2}{m-1}+s^2}\nu - \mathrm{x}_1\right)\\
&= \mathrm{x}_1 + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\frac{\frac{m-1}{\sigma^2+\sigma_o^2}+\frac{1}{s^2}}{\frac{m}{\sigma^2+\sigma_o^2}+\frac{1}{s^2}}\left(\frac{\sum_{i=2}^m x_i}{m-1} - \mathrm{x}_1 + \frac{\frac{\sigma^2+\sigma_o^2}{m-1}}{\frac{\sigma^2+\sigma_o^2}{m-1}+s^2}(\nu - \frac{\sum_{i=2}^m x_i}{m-1})\right)\\
&= \mathrm{x}_1 + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\frac{1+\frac{\sigma^2+\sigma_o^2}{(m-1)s^2}}{1+\frac{\sigma^2+\sigma_o^2}{ms^2}}\left(\frac{\sum_{i=1}^m x_i}{m} - \mathrm{x}_1 + \frac{m-1}{m}\frac{\frac{\sigma^2+\sigma_o^2}{m-1}}{\frac{\sigma^2+\sigma_o^2}{m-1}+s^2}(\nu - \frac{\sum_{i=2}^m x_i}{m-1})\right)\\
&= \mathrm{x}_1 + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\frac{1+\frac{\sigma^2+\sigma_o^2}{(m-1)s^2}}{1+\frac{\sigma^2+\sigma_o^2}{ms^2}}\left(\frac{\sum_{i=1}^m \mathrm{x}_i}{m} - \mathrm{x}_1\right) + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\frac{\frac{\sigma^2+\sigma_o^2}{m}}{s^2+\frac{\sigma^2+\sigma_o^2}{m}}\left(\nu - \frac{\sum_{i=2}^m \mathrm{x}_i}{m-1}\right)\\
&= \mathrm{x}_1 + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\left(1+\frac{1}{m-1}\frac{\frac{\sigma^2+\sigma_o^2}{m}}{s^2+\frac{\sigma^2+\sigma_o^2}{m}}\right)\left(\frac{\sum_{i=1}^m \mathrm{x}_i}{m} - \mathrm{x}_1\right) + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\frac{\frac{\sigma^2+\sigma_o^2}{m}}{s^2+\frac{\sigma^2+\sigma_o^2}{m}}\left(\nu - \frac{\sum_{i=2}^m \mathrm{x}_i}{m-1}\right)\\
&= \mathrm{x}_1 + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\left(\frac{\sum_{i=1}^m \mathrm{x}_i}{m} - \mathrm{x}_1\right) + \frac{\sigma_o^2}{\sigma^2+\sigma_o^2}\frac{\frac{\sigma^2+\sigma_o^2}{m}}{s^2+\frac{\sigma^2+\sigma_o^2}{m}}\left(\nu - \frac{\sum_{i=1}^m \mathrm{x}_i}{m}\right). \qquad (9)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}_G[\|\theta_{1:m} - \hat{\theta}^{\mathrm{JS}}_{1:m}(\mathrm{x}_{1:m})\|^2] = \mathbb{E}_G\left[m\sigma_o^2 + \|\hat{\theta}_{1:m} - \mathrm{x}_{1:m}\|^2 + 2\sigma_o^2\sum_{i=1}^m \frac{\partial}{\partial \mathrm{x}_i}(\hat{\theta}_i^{\mathrm{JS}} - \mathrm{x}_i)\right] \qquad (11)\\
&= \mathbb{E}_G\left[m\sigma_o^2 + \frac{\sigma_o^4(m-3)^2}{\|\mathrm{x}_{1:m} - \mathbf{1}_m\bar{\mathrm{x}}\|^2} - 2\sigma_o^4(m-3)\sum_{i=1}^m \frac{\frac{m-1}{m}\|\mathrm{x}_{1:m} - \mathbf{1}_m\bar{\mathrm{x}}\|^2 - 2\frac{m-1}{m}(\mathrm{x}_i - \bar{\mathrm{x}})^2 + \frac{2(\mathrm{x}_i - \bar{\mathrm{x}})}{m}\sum_{j\neq i}(\mathrm{x}_j - \bar{\mathrm{x}})}{\|\mathrm{x}_{1:m} - \mathbf{1}_m\bar{\mathrm{x}}\|^4}\right]\\
&= \mathbb{E}_G\left[m\sigma_o^2 + \frac{\sigma_o^4(m-3)^2}{\|\mathrm{x}_{1:m} - \mathbf{1}_m\bar{\mathrm{x}}\|^2} - 2\sigma_o^4(m-3)\sum_{i=1}^m \frac{\frac{m-1}{m}\|\mathrm{x}_{1:m} - \mathbf{1}_m\bar{\mathrm{x}}\|^2 - 2\frac{1}{m}(\mathrm{x}_i - \bar{\mathrm{x}})^2}{\|\mathrm{x}_{1:m} - \mathbf{1}_m\bar{\mathrm{x}}\|^4}\right]\\
&= m\sigma_o^2 - \sigma_o^4(m-3)^2\mathbb{E}_G\left[\frac{1}{\|\mathrm{x}_{1:m} - \mathbf{1}_m\bar{\mathrm{x}}\|^2}\right] \quad \text{(The above analysis does not rely on the distribution of } \theta_{1:m}\text{)}\\
&= m\sigma_o^2 - \frac{\sigma_o^4(m-3)}{\sigma_o^2+\sigma^2} = \frac{m\sigma_o^2\sigma^2}{\sigma_o^2+\sigma^2} + \frac{3\sigma_o^4}{\sigma_o^2+\sigma^2}. \qquad (12)
\end{aligned}
$$

$$
\begin{aligned}
&+ \mathbb{E}\left[2\sigma_o^2\sum_{i=1}^m\sum_{k=1}^d \frac{\partial}{\partial \mathrm{x}_{i,d}}(\hat{\boldsymbol{\theta}}^{\mathrm{HJS},\sigma^2}_{i,k} - \mathrm{x}_{i,k})\right]\\
&= md\sigma_o^2 + \mathbb{E}\left[\frac{\sigma_o^4\sum_k\|\mathrm{x}_{1:m,k} - \bar{\mathrm{x}}_k\mathbf{1}_m\|^2}{(\sigma^2+\sigma_o^2)^2} + \frac{\sigma_o^4(d-3)^2}{m\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}_d\|^2}\right]\\
&\quad - 2\sigma_o^2\mathbb{E}\left[\frac{\sigma_o^2 d(m-1)}{\sigma^2+\sigma_o^2} + \frac{\sigma_o^2(d-3)^2}{m\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}_d\|^2}\right]\\
&= md\sigma_o^2\\
&\quad - \frac{\sigma_o^4(2d(m-1)(\sigma^2+\sigma_o^2) - \mathbb{E}[\sum_k\|\mathrm{x}_{1:m,k} - \bar{\mathrm{x}}_k\mathbf{1}_m\|^2])}{(\sigma^2+\sigma_o^2)^2}\\
&\quad - \frac{\sigma_o^4(d-3)^2}{m}\mathbb{E}\left[\frac{1}{\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}_d\|^2}\right]\\
&= md\frac{\sigma^2\sigma_o^2}{\sigma^2+\sigma_o^2} + d\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} - \frac{d-3}{m}\mathbb{E}\left[\frac{(d-3)\sigma_o^4}{\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}_d\|^2}\right].
\end{aligned}
$$

E. Omitted calculation in the proof of Theorem III.4

We analyze the MSE of the Hierarchical JS estimator with unknown variance $\hat{\boldsymbol{\theta}}^{\mathrm{HJS}}_{i,k}$ by Stein's unbiased risk estimator (SURE). Since

$$
\begin{aligned}
&\sum_{i,k}\frac{\partial}{\partial \mathrm{x}_{i,k}}\left(\frac{\sigma_o^2(d(m-1)-2)}{\sum_{h=1}^d\|\mathrm{x}_{1:m,h} - \bar{\mathrm{x}}_h\mathbf{1}_m\|^2}(\bar{\mathrm{x}}_k - \mathrm{x}_{i,k})\right)\\
&= -\sigma_o^2\frac{(d(m-1)-2)^2}{\sum_{h=1}^d\|\mathrm{x}_{1:m,h} - \bar{\mathrm{x}}_h\mathbf{1}_m\|^2},
\end{aligned}
$$

we have Eq (13) - Eq (15).

Since $L(\hat{\boldsymbol{\theta}}_G; G) = d\frac{\sigma^4}{\sigma^2+\sigma_o^2}$, it followed by Eq (14) that

$$
\begin{aligned}
&\mathrm{TotReg}(\hat{\boldsymbol{\theta}}^{\mathrm{HJS}}_{1:m}; \mathcal{N}(\mu,\sigma^2))\\
&= (d+2)\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} - \frac{d-3}{m}\mathbb{E}\left[\frac{(d-3)\sigma_o^4}{\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}\|^2}\right].
\end{aligned}
$$

$$L(\boldsymbol{\theta}_{1:m}, \hat{\boldsymbol{\theta}}^{\text{HJS}}_{1:m}) = \mathbb{E}\left[md\sigma_o^2 + \|\hat{\boldsymbol{\theta}}^{\text{HJS}}_{1:m,1:d} - \mathrm{x}_{1:m,1:d}\|^2 + 2\sigma_o^2 \sum_{i=1}^{m}\sum_{k=1}^{d} \frac{\partial}{\partial \mathrm{x}_{i,d}}(\hat{\boldsymbol{\theta}}^{\text{HJS}}_{i,k} - \mathrm{x}_{i,k})\right] \tag{13}$$

$$= md\sigma_o^2 + \mathbb{E}\left[\frac{\sigma_o^4(d(m-1)-2)^2}{\sum_{k=1}^{d}\|\mathrm{x}_{1:m,k} - \bar{\mathrm{x}}_k \mathbf{1}_m\|^2} + \frac{\sigma_o^4(d-3)^2}{m^2\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}_d\|^2}\right] - 2\sigma_o^2\mathbb{E}\left[\frac{\sigma_o^2(d(m-1)-2)^2}{\sum_{k=1}^{d}\|\mathrm{x}_{1:m,k} - \bar{\mathrm{x}}_k \mathbf{1}_m\|^2} + \frac{\sigma_o^2(d-3)^2}{m\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}_d\|^2}\right]$$

$$= md\sigma_o^2 - \sigma_o^2\mathbb{E}\left[\frac{\sigma_o^2(d(m-1)-2)^2}{\sum_{k=1}^{d}\|\mathrm{x}_{1:m,k} - \bar{\mathrm{x}}_k \mathbf{1}_m\|^2} + \frac{\sigma_o^2(d-3)^2}{m\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}_d\|^2}\right]$$

$$= md\sigma_o^2 - md\frac{\sigma^4}{\sigma^2+\sigma_o^2} + d\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} + 2\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} - \frac{d-3}{m}\mathbb{E}\left[\frac{(d-3)\sigma_o^4}{\|\bar{\mathrm{x}}_{1:d} - \bar{\mathrm{x}}\mathbf{1}_d\|^2}\right] \tag{14}$$

$$= md\sigma_o^2 - md\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} + d\frac{\sigma_o^4}{\sigma^2+\sigma_o^2}\left(1 - \frac{\sigma^2+\sigma_o^2}{ms^2+\sigma^2+\sigma_o^2}\right) + 2\frac{\sigma_o^4}{\sigma^2+\sigma_o^2} + 3\frac{\sigma_o^4}{ms^2+\sigma^2+\sigma_o^2} \tag{15}$$

Further discussion. Note that it has been verified that the James-Stein estimator outperforms the maximum likelihood estimator in terms of smaller MSE, which indicates that collaboration is always better for this problem. The standard James-Stein estimator in this vector case is

$$\hat{\boldsymbol{\theta}}^{\text{JS}}_{i,k} = \mathrm{x}_{i,k} + \frac{\sigma_o^2(dm-3)}{\sum_{j,h}(\mathrm{x}_{j,h} - \bar{\mathrm{x}})^2}(\bar{\mathrm{x}} - \mathrm{x}_{i,k}),$$

with MSE

$$L(\hat{\boldsymbol{\theta}}^{\text{JS}}_{1:m}; G) = md\sigma_o^2 - \mathbb{E}_G\left[\frac{\sigma_o^4(dm-3)^2}{\sum_{j,h}(\mathrm{x}_{j,h} - \bar{\mathrm{x}})^2}\right]$$

$$= md\sigma_o^2 - \mathbb{E}_G\left[\frac{\sigma_o^4(dm-3)^2}{\sum_{j,h}(\mathrm{x}_{j,h} - \bar{\mathrm{x}}_h)^2 + \sum_h(\bar{\mathrm{x}}_h - \bar{\mathrm{x}})^2}\right].$$

F. Omitted proofs in communication constrained estimation

Proof of Theorem IV.4. We first create a class of distributions $\mathcal{G} = \{\mathcal{N}(\mu_v, \sigma^2\mathbb{I}_d) : v \in \mathcal{V}\}$ by auxiliary vector $v \in \mathcal{V} = \{\pm 1\}^d$, where $\mu_v = \delta v$.

For any estimator $\hat{\boldsymbol{\theta}}_{1:m}(\mathbf{x}_i, Z_{-i})$, let $\hat{G}(\mathbf{x}_i, Z_{-i}) = \arg\min_{G\in\mathcal{G}} \|\hat{\boldsymbol{\theta}}_i(\mathbf{x}_i, Z_{-i}) - \hat{\boldsymbol{\theta}}_G(\mathbf{x}_i)\|$ be a prior distribution estimator. We know

$$2\|\hat{\boldsymbol{\theta}}_i(\mathbf{x}_i, Z_{-i}) - \hat{\boldsymbol{\theta}}_G(\mathbf{x}_i)\|$$
$$\geq \|\hat{\boldsymbol{\theta}}_i(\mathbf{x}_i, Z_{-i}) - \hat{\boldsymbol{\theta}}_{\hat{G}(\mathbf{x}_i, Z_{-i})}(\mathbf{x}_i)\| + \|\hat{\boldsymbol{\theta}}_i(\mathbf{x}_i, Z_{-i}) - \hat{\boldsymbol{\theta}}_G(\mathbf{x}_i)\|$$
$$\geq \|\hat{\boldsymbol{\theta}}_{\hat{G}(\mathbf{x}_i, Z_{-i})}(\mathbf{x}_i) - \hat{\boldsymbol{\theta}}_G(\mathbf{x}_i)\|,$$

and it follows that

$$\text{TotReg}(\hat{\boldsymbol{\theta}}_{1:m}; \mathcal{G}) = \sup_{G\in\mathcal{G}} \mathbb{E}_G\left[\sum_{i=1}^{m}\|\hat{\boldsymbol{\theta}}_i(\mathbf{x}_i, Z_{-i}) - \hat{\boldsymbol{\theta}}_G(\mathbf{x}_i)\|^2\right]$$
$$\geq \frac{1}{4}\sup_{v\in\mathcal{V}} \mathbb{E}_G\left[\sum_{i=1}^{m}\|\hat{\boldsymbol{\theta}}_{\hat{G}(\mathbf{x}_i, Z_{-i})}(\mathbf{x}_i) - \hat{\boldsymbol{\theta}}_G(\mathbf{x}_i)\|^2\right].$$

Let $\hat{V}(\mathbf{x}_i, Z_{-i})$ be the corresponding auxiliary vector of $\hat{G}(\mathbf{x}_i, Z_{-i})$, and since $\hat{\boldsymbol{\theta}}_G(\boldsymbol{x}) = \boldsymbol{x} + \frac{\sigma_o^2/n}{\sigma^2+\sigma_o^2/n}(\boldsymbol{\mu} - \boldsymbol{x})$, we know

$$\text{TotReg}(\hat{\boldsymbol{\theta}}_{1:m}; \mathcal{G})$$
$$\geq \frac{1}{4}\sup_{v\in\mathcal{V}} \mathbb{E}_{\boldsymbol{\mu}_v}\left[\left(\frac{\sigma_o^2/n}{\sigma_o^2/n+\sigma^2}\right)^2 \sum_{i=1}^{m}\|\boldsymbol{\mu}_{\hat{V}(\mathbf{x}_i, Z_{-i})} - \boldsymbol{\mu}_v\|^2\right]$$
$$\geq \max_{v\in\mathcal{V}} \mathbb{E}_{\boldsymbol{\mu}_v}\left[\delta^2\left(\frac{\sigma_o^2/n}{\sigma_o^2/n+\sigma^2}\right)^2 \sum_{i=1}^{m} d_{\text{ham}}(\hat{V}(\mathbf{x}_i, Z_{-i}), v)\right],$$

where d_{ham} is hamming distance. Let V uniformly distributed over $\mathcal{V}$, we then have

$$\text{TotReg}(\hat{\boldsymbol{\theta}}_{1:m}; \mathcal{G})$$
$$\geq \delta^2\left(\frac{\sigma_o^2/n}{\sigma_o^2/n+\sigma^2}\right)^2 \sum_{i=1}^{m}\mathbb{E}[d_{\text{ham}}(\hat{V}(\mathbf{x}_i, Z_{-i}), V)]$$
$$\geq \delta^2\left(\frac{\sigma_o^2/n}{\sigma_o^2/n+\sigma^2}\right)^2 (\lfloor t\rfloor + 1)\sum_{i=1}^{m}\mathbb{P}(d_{\text{ham}}(\hat{V}(\mathbf{x}_i, Z_{-i}), V) > t),$$

where the first inequality is by replacing the $\max_{v\in\mathcal{V}}$ by expectation and the second inequality is by Markov inequality. By distance-based Fano's inequality [1], [8] that $\mathbb{P}(d_{\text{ham}}(\hat{V}(\mathbf{x}_i, Z_{-i}), V) > t) \geq 1 - \frac{I(V;\mathbf{x}_i, Z_{-i}) + \log 2}{\log \frac{|\mathcal{V}|}{\max_v |\{v' : d_{\text{ham}}(v,v') \leq t\}|}}$ and taking $t = d/6$ gives

$$\text{TotReg}(\hat{\boldsymbol{\theta}}_{1:m}; \mathcal{G}) \geq \sum_{i=1}^{m}\delta^2\left(\frac{\sigma_o^2/n}{\sigma_o^2/n+\sigma^2}\right)^2 (\lfloor d/6\rfloor + 1)\left(1 - \frac{I(V;\mathbf{x}_i, Z_{-i}) + \log 2}{d/6}\right).$$

Since $\mathbf{x}_i, \{Z_j\}_{j\neq i}$ are conditionally independent given V, we have $I(V;\mathbf{x}_i, Z_{-i}) \leq I(V;\mathbf{x}_i) + \sum_{j\neq i} I(V;Z_j)$.

Note that $\mathbf{x}_i | V = v \sim \mathrm{P}_v := \mathcal{N}(\boldsymbol{\mu}_v, (\sigma^2 + \sigma_o^2/n)\mathbb{I}_d)$. Let V' be an independent copy of V, we have

$$I(V;\mathbf{x}_i) \leq \frac{1}{2^{2d}}\sum_{v,v'}\text{KL}(\mathrm{P}_v \| \mathrm{P}_{v'})$$
$$\leq \frac{\delta^2}{2(\sigma_o^2/n+\sigma^2)}\mathbb{E}\left[d_{\text{ham}}(V, V')\right] = \frac{\delta^2 d}{4(\sigma_o^2/n+\sigma^2)}.$$

It requires $\delta^2 \leq \frac{\sigma_o^2/n+\sigma^2}{100}$ to imply $I(V;\mathbf{x}_i) \leq d/400$. $I(V;Z_j)$ is then bounded by strong data processing inequality as in [8, Lemma 5], which requires

$$\delta^2 \leq \min\left\{1, \frac{\sigma_o^2/n+\sigma^2}{400\log(m)}, \frac{d(\sigma_o^2/n+\sigma^2)}{100\sum_{j\neq i}\min(25\cdot 128 B_j \log m, d)}\right\},$$

to make $\sum_{j\neq i} I(V;Z_j) \leq d/40$. When δ satisfies all the above requirements, $\left(1-\frac{I(V;\mathbf{x}_i,Z_{-i})+\log 2}{d/6}\right) \geq \frac{1}{2}$ and the lower bound is thus concluded. $\square$