Iterative Bayesian Learning for Crowdsourced Regression (Supplementary Material)

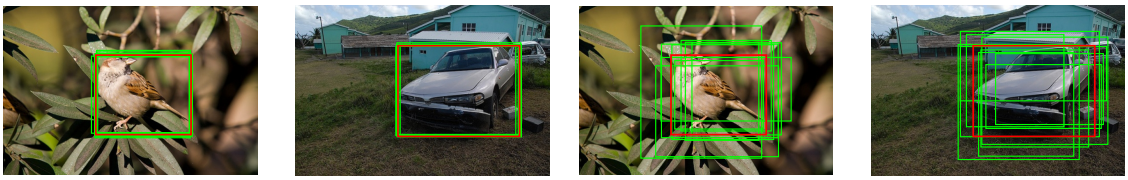
Jungseul Ok* **Sewoong Oh*** **Yunhun Jang†** **Jinwoo Shin†** **Yung Yi†**
*University of Washington, †Korea Advanced Institute of Science and Technology

A MOTIVATING EXPERIMENT: IMPORTANCE OF ESTIMATOR

Crowdsourcing is a primary marketplace to get labels on training datasets, to be used to train machine learning models. In this section, using both semi-synthetic and real datasets, we investigate the impact of having higher quality labels on real-world machine learning tasks. We show that sophisticated regression algorithms like BI can produce high quality labels on the crowdsourced training datasets, improving the end-to-end performance of convolutional neural network (CNN) on visual object detection or human age prediction. This highlights the importance of estimator but also justifies the use of the proposed BI, in a real world system.

A.1 Visual Object Detection

Emulating a crowdsourcing system. To do so, we use PASCAL visual object classes (VOC) datasets from (Everingham et al., 2015): VOC-07/12 consisting of 40,058 annotated objects in 16,551 images. Each object is annotated by a bounding box expressed by two opposite corner points. We emulate the crowdsourcing system with a random $(\ell = 3, r = 10)$-regular bipartite graph where each image is assigned to 3 workers and each worker is assigned 10 images ($\simeq 24.2$ objects on average) to draw the bounding boxes of every object in the assigned images. Each worker has variance drawn uniformly at random from support $\mathcal{S} = \{10, 1000\}$. and generates noisy responses of which examples are shown in Figure S1.



(a) When $\sigma_u^2 = 10$ (b) When $\sigma_u^2 = 1000$

Figure S1: Examples of object annotations by a worker u with $\sigma_u^2 = 10$ or 1000.

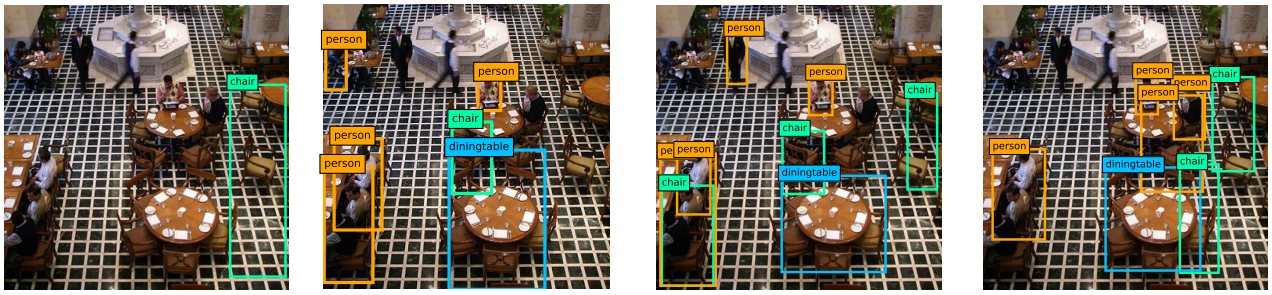
Evaluation on visual object detection task. Using each training dataset from four different estimators (Average, NBI, BI, Strong-Oracle), we train[1] CNN of single shot multibox detector (SSD) model (Liu et al., 2016), which shows the state-of-the-art performance. Then we evaluate the trained SSD's in terms of the mean average precision (mAP) which is a popular benchmarking metric for the datasets (see Table S1). Intuitively, a high mAP means more true positive and less false positive detections.

Comparing mAP of Average, mAP's of BI and NBI are 4% mAP higher as Figure S2 also visually shows the improvement. Note that achieving a similar amount of improvement is highly challenging. Indeed, Faster-RCNN in (Ren et al., 2015) is proposed to improve the mAP of Fast-RCNN in (Girshick, 2015) from 70.0% to 73.2%. Later, SSD in (Liu et al., 2016) is proposed to achieve 4% mAP improvement over Faster-RCNN.

[1]As suggested by (Liu et al., 2016), we train SSD using 120,000 iterations where the learning rate is initialized at 4×10^{-5}, and is decreased by factor 0.1 at 80,000-th and 100,000-th iterations.

Table S1: Estimation quality of Average, NBI, BI, and Strong-Oracle on crowdsourced VOC-07/12 datasets from virtual workers in terms of MSE, and performance of SSD's trained with the estimated dataset and ground truth (VOC-07/12) in terms of mean average precision (mAP); mean portion of the output bounding box overlapped on the ground truth (Overlap).

Estimator	Data noise (MSE)	Testing accuracy (mAP)	(Overlap)
Average	355.6	71.80	0.741
NBI	116.1	75.62	0.767
BI	109.8	75.94	0.772
Strong-Oracle	109.8	76.05	0.774
Ground truth	-	77.79	0.784



(a) Average (b) NBI (c) BI (d) Strong-Oracle

Figure S2: Examples of detections of SSD trained by the crowdsourced VOC-07/12 datasets by Average, NBI, BI, and Strong-Oracle.

A.2 Human Age Prediction

Real-world dataset. We also perform similar experiment using datasets from a *real-world* crowdsourcing system. We use FG-NET datasets which has been widely used as a benchmark dataset for facial age estimation (Lanitis, 2008). The dataset contains 1,002 photos of 82 individuals' faces, in which each photo has biological age as ground truth. Furthermore, (Han et al., 2015) provide crowdsourced labels on FG-NET datasets, in which 165 workers in Amazon Mechanical Turk (MTurk) answer their own age estimation on given subset of 1,002 photos so that each photo has 10 answers from workers, while each worker provides a different number (from 1 to 457) of answers, and 60.73 answers in average.

In the dataset, we often observe two extreme classes of answers for a task: a few outliers and consensus among majority. For example, for Figures S3(a) and S3(b), there exist noisy answers 5 and 7, respectively, which are far from majority 1 and 55, respectively. Such observations suggest to choose a simple support, e.g. $\mathcal{S} = \{\sigma^2_{\text{good}}, \sigma^2_{\text{bad}}\}$. In particular, without any use of ground truth, we first run NBI and use the top 10% and bottom 10% workers' reliabilities as the binary support, which is $\mathcal{S}_{\text{est}} = \{6.687, 62.56\}$ in our experiment.

Evaluation on human age prediction task. We first compare the estimation of BI to other algorithms as reported in Table S2. Observe that MSE of BI with the binary support $\mathcal{S}_{\text{est}}$ is close to that of Strong-Oracle, while the other algorithms have some gaps. This result from real workers supports the idea of simplified workers' noise level in our model. We also evaluate the impact on de-noising process for human age prediction. To this end, using the pruned datasets from different estimators, we train[2] one of the state-of-art CNN models, called VGG-16 (Simonyan and Zisserman, 2014), under some modification proposed by (Rothe et al., 2015) for human age prediction. Although the crowdsourced dataset FG-NET is not large-scale in order to see performance

[2] We train VGG-16 using batch normalization with standard hyper parameter setting, where we initialize based on the imagenet pre-trained model. To regressed the estimated age of the given face images, we replaced final layer of VGG-16 with one dimensional linear output layer, and fine-tuned all the layers with initial learning rate 0.01 (and divided by 10 after 30, 60, 90 epoch). Protocol of measuring model performance is standard Leave One Person Out (LOPO) which uses images of 81 subjects for training and use remaining subject for test, and the final result is averaged over the total 82 model training (Panis et al., 2016).

Table S2: Estimation quality of Average, NBI, BI with $\mathcal{S}_{\text{est}}$, and Strong-Oracle on crowdsourced FG-NET datasets from Amazon MTurk workers in terms of MSE, and performance of VGG-16's trained with the estimated datasets and the ground truth dataset (FG-NET) in terms of median absolute error (MDAE).

Estimator	Data noise (MSE)	Testing error (MDAE)
Average	34.99	3.227
NBI	32.80	3.135
BI	28.72	3.100
Strong-Oracle	28.45	3.003
Ground truth	-	1.822

(a) 1-year-old (b) 35-year-old

Figure S3: Easy and hard samples from FG-NET in terms of average absolute error of crowd workers' answers: (1,1,1,1,1,1,1,1,2,5) and (7,51,52,55,55,55,59,63,66,67) on photo of (a) 1-year-old, and (b) 35-year-old, resp.

difference, models trained by both BI and NBI show superiority to that of Average (which is widely used in practical crowdsourcing systems), in terms of median absolute errors (MDAE), as reported in Table S2.

B MODEL DERIVATIONS

B.1 Calculation of $\bar{\mu}$

We first show that the posterior density of μ_i given $A_i = y_i := \{y_{iu} \in \mathbb{R}^d : u \in M_i\}$ and $\sigma^2_{M_i}$ is a Gaussian density in the following:

$$f_{\mu_i}[x \,|\, A_i = y_i, \sigma^2_{M_i}] = \frac{f_{\mu_i}[x]\, f_{A_i}\left[y_i \,|\, \mu_i = x_i, \sigma^2_{M_i}\right]}{f_{A_i}\left[y_i \,|\, \sigma^2_{M_i}\right]} \tag{S1}$$

$$= \phi\left(x \,|\, \bar{\mu}_i\left(y_i, \sigma^2_{M_i}\right), \bar{\sigma}^2_i\left(\sigma^2_{M_i}\right)\right) \tag{S2}$$

where we define $\bar{\sigma}^2_i : \mathcal{S}^{M_i} \to \mathbb{R}$ and $\bar{\mu}_i : \mathbb{R}^{d\times M_i} \times \mathcal{S}^{M_i} \to \mathbb{R}^d$ as follows

$$\bar{\sigma}^2_i\left(\sigma^2_{M_i}\right) := \frac{1}{\frac{1}{\tau^2} + \sum_{u\in M_i} \frac{1}{\sigma^2_u}}, \quad \text{and} \quad \bar{\mu}_i\left(A_i, \sigma^2_{M_i}\right) := \bar{\sigma}^2_i\left(\sigma^2_{M_i}\right)\left(\frac{\nu_i}{\tau^2} + \sum_{u\in M_i} \frac{A_{iu}}{\sigma^2_u}\right).$$

The Gaussian posterior density (S2) follows from:

$$f_{\mu_i}[x] f_{A_i}\left[y_i \,\middle|\, \mu_i = x, \sigma^2_{M_i}\right] = \phi(x \,|\, \nu_i, \tau^2) \prod_{u\in M_i} \phi(y_{iu} \,|\, \mu_i, \sigma^2_u)$$

$$= \mathcal{C}_i\left(y_i, \sigma^2_{M_i}\right) \phi\left(x \,|\, \bar{\mu}_i\left(y_i, \sigma^2_{M_i}\right), \bar{\sigma}^2_i\left(\sigma^2_{M_i}\right)\right)$$

where we have $f_{A_i}[y_i \,|\, \sigma^2_{M_i}] = \mathcal{C}_i(y_i, \sigma^2_{M_i})$ with

$$\mathcal{C}_i(A_i, \sigma^2_{M_i}) := \left(\frac{2\pi\bar{\sigma}^2_i(\sigma^2_{M_i})}{2\pi\tau_i^2 \prod_{u\in M_i}(2\pi\sigma^2_u)}\right)^{\frac{d}{2}} e^{-\mathcal{D}_i(A_i, \sigma^2_{M_i})}, \text{ and}$$

$$\mathcal{D}_i(A_i, \sigma^2_{M_i}) := \frac{1}{2}\bar{\sigma}^2_i(\sigma^2_{M_i})\left(\sum_{u\in M_i} \frac{\|A_{iu} - \nu_i\|_2^2}{\sigma^2_u \tau^2} + \sum_{v\subset M_i\setminus\{u\}} \frac{\|A_{iu} - A_{iv}\|_2^2}{\sigma^2_u \sigma^2_v}\right).$$

The Gaussian density in (S2) leads to the posterior mean, which is weighted average of the prior mean and the worker responses, each weighted by the inverse of its variance:

$$\mathbb{E}[\mu_i \,|\, A_i, \sigma^2_{M_i}] = \bar{\mu}_i(A_i, \sigma^2_{M_i}).$$

Thus, the optimal estimator $\hat{\mu}^*_i(A)$ is given as (2).

B.2 Factorization of Joint Probability

Using Bayes' theorem, it is not hard to write the joint probability of σ^2 given $A = y = \{y_{iu} \in \mathbb{R}^d : (i,u) \in E\}$,

$$\mathbb{P}[\sigma^2 \,|\, A = y] \propto f_A[y \,|\, \sigma^2] = \prod_{i \in V} f_{A_i}[y_i \,|\, \sigma^2_{M_i}] = \prod_{i \in V} \mathcal{C}_i(y_i, \sigma^2_{M_i}) \,.$$

C PROOFS OF LEMMAS

C.1 Proof of Lemma 1

We first introduce an inference problem and connect its error rate to the expectation of likelihood of worker ρ's σ^2_ρ given A. Let $s_\rho \in \{1, \dots, S\}$ be the index of $\tilde{\sigma}^2_\rho$, i.e., $\tilde{\sigma}^2_\rho = \sigma^2_{s_\rho}$. Consider the classification problem recovering given but latent s from $A_{\rho,2k}$, where $A_{\rho,2k}$ is generated from the crowdsourced regression model with fixed but hidden $\sigma^2 = \tilde{\sigma}^2$. More formally, the problem is formulated as the following optimization problem:

$$\underset{\hat{s}_\rho : \text{estimator}}{\text{minimize}} \ \mathbb{P}[s_\rho \neq \hat{s}_\rho(A_{\rho,2k})] \tag{S3}$$

where the optimal estimator, denoted by $\hat{s}^*_\rho$, minimizes the classification error rate. From the standard Bayesian argument, the optimal estimator $\hat{s}^*_\rho$ is given $A_{\rho,2k}$ as

$$\hat{s}^*_\rho(A_{\rho,2k}) := \underset{s'_\rho = 1,\dots,S}{\arg\max} \ \mathbb{P}[s_\rho = s'_\rho \,|\, A_{\rho,2k}] \,. \tag{S4}$$

From the construction of the optimal estimator in (S4), it is not hard to check

$$\mathbb{P}_{\tilde{\sigma}^2}[s_\rho = \hat{s}^*_\rho(A_{\rho,2k})] \ := \ \mathbb{P}[s_\rho = \hat{s}^*_\rho(A_{\rho,2k}) \,|\, \sigma^2 = \tilde{\sigma}^2] \ \leq \ \mathbb{E}_{\tilde{\sigma}^2}\left[\mathbb{P}[\sigma^2_\rho = \tilde{\sigma}^2_\rho \,|\, A_{\rho,2k}]\right] \,. \tag{S5}$$

Thus an upper bound of the average error rate of an estimator for (S3) will provide an upper bound of $\mathbb{E}_{\tilde{\sigma}^2}\left[\mathbb{P}[\sigma^2_\rho \neq \tilde{\sigma}^2_\rho \,|\, A_{\rho,2k}]\right]$ since the optimal estimator minimizes the average error rate. Indeed, we have

$$\begin{aligned}
\mathbb{E}_{\tilde{\sigma}^2}\left[\mathbb{P}[\sigma^2_\rho \neq \tilde{\sigma}^2_\rho \,|\, A_{\rho,2k}]\right] &\leq 1 - \mathbb{E}_{\tilde{\sigma}^2}\left[\mathbb{P}[\sigma^2_\rho = \tilde{\sigma}^2_\rho \,|\, A_{\rho,2k}]\right] \\
&= \mathbb{E}[\mathbb{P}_{\tilde{\sigma}^2}[s_\rho \neq \hat{s}^*_\rho(A_{\rho,2k})]] \\
&= \min_{\hat{s}_\rho : \text{estimator}} \mathbb{P}[s_\rho \neq \hat{s}_\rho(A_\rho, 2k)] \,.
\end{aligned}$$

Consider a simple estimator for (S3), denoted by $\hat{s}^\dagger_\rho$, which uses only $A_{\rho,2} \subset A_{\rho,2k}$ as follows:

$$\hat{s}^\dagger_\rho(A_{\rho,2}) = \underset{s'_\rho = 1,\dots,S}{\arg\min} \left| (\sigma^2_{s'_\rho} + \sigma^2_{\mathsf{avg}}(\mathcal{S})) - \hat{\sigma}^2(A_{\rho,2k}) \right| \tag{S6}$$

where we define

$$\sigma^2_{\mathsf{avg}}(\mathcal{S}) := \frac{\sum_{s'=1,\dots,S} \sigma^2_{s'}}{S(\ell - 1)} \,, \hat{\sigma}^2(A_{\rho,2}) := \frac{1}{r} \sum_{i \in N_\rho} \hat{\sigma}^2_i(A_i) \,, \text{and } \hat{\sigma}^2_i(A_i) := \left\| \frac{\sum_{u \in M_i \setminus \{\rho\}} A_{iu}}{\ell - 1} - A_{i\rho} \right\|_2^2 \,.$$

From now on, we condition $\sigma^2_{\partial^2 \rho}$ additionally to σ^2_ρ where $\partial^2 \rho$ is the set of ρ's grandchildren in $G_{\rho,2}$. For every $i \in N_\rho$, we define

$$a_i := \sum_{u \in M_i \setminus \{\rho\}} \frac{\tilde{\sigma}^2_u}{(\ell - 1)^2} + \tilde{\sigma}^2_\rho \,, \text{ and } Z_i := \frac{\sum_{u \in M_i \setminus \{\rho\}} A_{iu}}{\ell - 1} - A_{i\rho} \,.$$

Since the conditional density of Z_i given $\sigma^2 = \tilde{\sigma}^2$ is $\phi(Z_i \,|\, 0, a_i)$, the conditional density of $\|Z_i\|_2^2 / a_i$ is χ^2-distribution with degree of freedom d. In addition, it is not hard to check that $\|Z_i\|_2^2$ is sub-exponential with parameters $((2a_i\sqrt{d})^2, 2a_i)$ such that for all $|\lambda| < \frac{1}{2a_i}$,

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\exp\left(\lambda\left(\|Z_i\|_2^2 - d a_i\right)\right)\right] = \left(\frac{e^{-a_i\lambda}}{\sqrt{1 - 2a_i\lambda}}\right)^d \leq \exp\left(\frac{(2a_i\sqrt{d})^2\lambda^2}{2}\right) \,.$$

Thus it follows that for all $|\lambda| \leq \min_{i \in N_\rho} \frac{1}{2a_i}$,

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\exp\left(\lambda \sum_{i \in N_\rho} \left(\|Z_i\|_2^2 - da_i\right)\right)\right] = \prod_{i \in N_\rho} \mathbb{E}_{\tilde{\sigma}^2}\left[\exp\left(\lambda\left(\|Z_i\|_2^2 - da_i\right)\right)\right] \leq \prod_{i \in N_\rho} \exp\left(\frac{(2a_i\sqrt{d})^2\lambda^2}{2}\right).$$

From this, it is straightforward to check that $r\hat{\sigma}^2(A_{\rho,2}) = \sum_{i \in N_\rho} \|Z_i\|_2^2$ is sub-exponential with parameters $((6\sigma_{\max}^2\sqrt{d})^2, 6\sigma_{\max}^2)$ since

$$0 \leq a_i \leq \sigma_{\max}^2\left(\frac{\ell+1}{\ell-1}\right) \leq 3\sigma_{\max}^2 \,. \tag{S7}$$

Using Bernstein bound, we have

$$\mathbb{P}_{\tilde{\sigma}^2}\left[\left|\hat{\sigma}^2(A_{\rho,2}) - \frac{\sum_{i \in N_\rho} a_i}{r}\right| \geq \frac{\varepsilon}{4}\right] \;\leq\; 2\exp\left(-\frac{\varepsilon r}{48\sigma_{\max}^2}\right) \tag{S8}$$

where we let $\mathbb{P}_{\tilde{\sigma}^2}$ denote the conditional probability given $\sigma^2 = \tilde{\sigma}^2$. Using Hoeffding bound with (S7), it follows that

$$\mathbb{P}_{\tilde{\sigma}^2}\left[\left|\frac{\sum_{i \in N_\rho} a_i}{r} - \left(\sigma_{\mathsf{avg}}^2(\mathcal{S}) + \sigma_\rho^2\right)\right| \geq \frac{\varepsilon}{4}\right] \;\leq\; 2\exp\left(-\frac{\varepsilon^2 r}{8\sigma_{\max}^2}\right) \,. \tag{S9}$$

Combining (S8) and (S9) and using the union bound, it follows that

$$\begin{aligned}\mathbb{P}_{\tilde{\sigma}^2}\left[s_\rho \neq \hat{s}_\rho^\dagger(A_{\rho,2})\right] &\leq \mathbb{P}_{\tilde{\sigma}^2}\left[\left|\hat{\sigma}^2(A_{\rho,2}) - \left(\sigma_{\mathsf{avg}}^2(\mathcal{S}) + \sigma_\rho^2\right)\right| > \frac{\varepsilon}{2}\right] \\ &\leq 2\left(\exp\left(-\frac{\varepsilon r}{48\sigma_{\max}^2}\right) + \exp\left(-\frac{\varepsilon^2 r}{8\sigma_{\max}^2}\right)\right) \\ &\leq 4\exp\left(-\frac{\varepsilon^2 r}{8(8\varepsilon+1)\sigma_{\max}^2}\right)\end{aligned} \tag{S10}$$

where for the first inequality we use $|\sigma_{s'}^2 - \sigma_{s''}^2| \geq \varepsilon$ for all $1 \leq s', s'' \leq S$ such that $s' \neq s'$. Hence, noting that $\hat{s}^\dagger$ cannot outperform the optimal one $\hat{s}^*$ in (S5), this performance guarantee on $\hat{s}^\dagger$ in (S10) completes the proof of Lemma 1.

C.2 Proof of Lemma 2

We begin with the underlying intuition on the proof. As Lemma 1 states, if there is the strictly positive gap $\varepsilon > 0$ between $\sigma_{\min}^2$ and $\sigma_{\max}^2$, one can recover $\sigma_\rho^2 \in \{\sigma_{\min}^2, \sigma_{\max}^2\}$ with small error using only the local information, i.e., $A_{\rho,2k}$. On the other hand, $A \setminus A_{\rho,2k}$ is far from ρ and is less useful on estimating σ_ρ^2. In the proof of Lemma 2, we quantify the decaying rate of information w.r.t. k.

We first introduce several notations for convenience. For $u \in W_{\rho,2k}$, let $T_u = (V_u, W_u, E_u)$ be the subtree rooted from u including all the offsprings of u in tree $G_{\rho,2k}$. Note that $T_\rho = G_{\rho,2k}$. We let $\partial W_u \subset W_{\rho,2k}$ denote the subset of worker on the leaves in T_u and let $A_u := \{A_{iv} : (i,v) \in E_u\}$. Since each worker u's σ_u^2 is a binary random variable, we define a function $s_u : \mathcal{S} \to \{+1, -1\}$ for the given $\tilde{\sigma}^2$ as follows:

$$s_u(\sigma_u^2) = \begin{cases} +1 & \text{if } \sigma_u^2 = \tilde{\sigma}_u^2 \\ -1 & \text{if } \sigma_u^2 \neq \tilde{\sigma}_u^2 \,. \end{cases}$$

It is enough to show

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\left|\mathbb{P}[s_\rho(\sigma_\rho^2) = +1 \,|\, A_{\rho,2k}, \sigma_{\partial W_\rho}^2] - \mathbb{P}[s_\rho(\sigma_\rho^2) = +1 \,|\, A_{\rho,2k}]\right|\right] \;\leq\; 2^{-k} \tag{S11}$$

since for each $u \in W$, $\mathbb{P}[\sigma_u^2 = \sigma_1^2] = \mathbb{P}[\sigma_u^2 = \sigma_2^2] = \frac{1}{2}$.

To do so, we first define

$$X_u := 2\mathbb{P}[s_u(\sigma_u^2) = +1 \,|\, A_u] - 1 \,, \quad \text{and} \quad Y_u := 2\mathbb{P}[s_u(\sigma_u^2) = +1 \,|\, A_u, , \sigma^2_{\partial W_\rho}] - 1$$

so that we have

$$\left|\mathbb{P}[s_\rho(\sigma_\rho^2) = +1 \,|\, A_{\rho,2k}, \sigma^2_{\partial W_\rho}] - \mathbb{P}[s_\rho(\sigma_\rho^2) = +1 \,|\, A_{\rho,2k}]\right| \;=\; \frac{1}{2}\,|X_\rho - Y_\rho| \,.$$

Using the above definitions of X_u and Y_u and noting $|X_u - Y_u| \le 2$, it is enough to show that for given non-leaf worker $u \in W_\rho \setminus \partial W_\rho$,

$$\mathbb{E}_{\tilde\sigma^2}\left[\left|X_u - Y_u\right|\right] \le \frac{1}{2|\partial^2 u|} \sum_{v \in \partial^2 u} \mathbb{E}_{\tilde\sigma^2}\left[\left|X_v - Y_v\right|\right] \tag{S12}$$

where we let $\partial^2 u$ denote the set of grandchildren of u in T_u.

To do so, we study certain recursions describing relations among X and Y. For notational convenience, we define g^+_{iu} and g^-_{iu} as follows:

$$g^+_{iu}(X_{\partial_u i}; A_i) := \sum_{\sigma'^2_{M_i} \in \mathcal{S}^{M_i} : \sigma'^2_u = \tilde\sigma^2_u} \mathcal{C}_i(A_i, \sigma'^2_{M_i}) \prod_{v \in \partial_u i} \frac{1 + s_v(\sigma'^2_v) X_v}{2}$$

$$g^-_{iu}(X_{\partial_i u}; A_i) := \sum_{\sigma'^2_{M_i} \in \mathcal{S}^{M_i} : \sigma'^2_u \neq \tilde\sigma^2_u} \mathcal{C}_i(A_i, \sigma'^2_{M_i}) \prod_{v \in \partial_u i} \frac{1 + s_v(\sigma'^2_v) X_v}{2} \,.$$

where we may omit A_i in the argument of g^+_{iu} and g^-_{iu} if A_i is clear from the context. Recalling the factor form of the joint probability of σ^2, i.e., and using Bayes' theorem with the fact that $\mathbb{P}[s_u(\sigma_u^2) = +1 \,|\, A_u] = \frac{1+X_u}{2}$ and some calculus, it is not hard to check

$$g^+_{iu}(X_{\partial_u i}; A_i) \propto \mathbb{P}\left[s_u(\sigma_u^2) = +1 \,|\, A_i, X_{\partial_u i}\right] \tag{S13}$$

$$g^-_{iu}(X_{\partial_i u}; A_i) \propto \mathbb{P}\left[s_u(\sigma_u^2) = -1 \,|\, A_i, X_{\partial_u i}\right] \,. \tag{S14}$$

From the above, it is straightforward to check that

$$\begin{aligned} X_u &= h_u(X_{\partial^2 u}) \\ &:= \frac{\prod_{i \in \partial u} g^+_{iu}(X_{\partial_u i}) - \prod_{i \in \partial u} g^-_{iu}(X_{\partial_u i})}{\prod_{i \in \partial u} g^+_{iu}(X_{\partial_u i}) + \prod_{i \in \partial u} g^-_{iu}(X_{\partial_u i})} \end{aligned} \tag{S15}$$

where we let ∂u be the task set of all the children of worker u and $\partial_u i$ be the worker set of all the children of i in tree T_u. Similarly, we also have

$$Y_u = h_u(Y_{\partial^2 u}) \,.$$

For simplicity, we now pick an arbitrary worker $u \in W_\rho$ which is neither the root nor a leaf, i.e., $u \notin \partial W_\rho$ and $u \neq \rho$, so that $\left|\partial^2 u\right| = (\ell - 1)(r - 1)$. It is enough to show (S12) for only u. To do so, we will use the mean value theorem. We first obtain a bound on the gradient of $h_u(x)$ for $x \in [-1, 1]^{\partial^2 u}$. Define $g^+_u(x) := \prod_{i \in \partial u} g^+_{iu}(x_{\partial_u i})$ and $g^-_u(x) := \prod_{i \in \partial u} g^-_{iu}(x_{\partial_u i})$. Using basic calculus, we obtain that for $v \in \partial_u i$,

$$\begin{aligned} \frac{\partial h_u}{\partial x_v} &= \frac{\partial}{\partial x_v} \frac{g^+_u - g^-_u}{g^+_u + g^-_u} \\ &= \frac{2}{(g^+_u + g^-_u)^2} \left(g^-_u \frac{\partial g^+_u}{\partial x_v} - g^+_u \frac{\partial g^-_u}{\partial x_v} \right) \\ &= \frac{2 g^+_u g^-_u}{(g^+_u + g^-_u)^2} \left(\frac{1}{g^+_{iu}} \frac{\partial g^+_{iu}}{\partial x_v} - \frac{1}{g^-_{iu}} \frac{\partial g^-_{iu}}{\partial x_v} \right) . \end{aligned}$$

Using the fact that for $x \in [-1,1]^{\partial^2 u}$, both g_u^+ and g_u^- are positive, it is not hard to show that

$$\frac{g_u^+ g_u^-}{(g_u^+ + g_u^-)^2} \le \sqrt{\frac{g_u^-}{g_u^+}}. \tag{S16}$$

We note here that one can replace g_u^-/g_u^+ with g_u^+/g_u^- in the upper bound. However, in our analysis, we use (S16) since we will take the conditional expectation $\mathbb{E}_{\tilde{\sigma}^2}$ which takes the randomness of A generated by the condition $\sigma^2 = \tilde{\sigma}^2$. Hence X_u and Y_u will be closer to 1 than -1 thus g_u^-/g_u^+ will be a tighter upper bound than g_u^+/g_u^-.

From (S16), it follows that for $x \in [-1,1]^{\partial^2 u}$ and $v \in \partial_u i$,

$$\left|\frac{\partial h_u}{\partial x_v}(x)\right| \le |g'_{uv}(x_{\partial_u i})| \prod_{j \in \partial u\,:\, j \neq i} \sqrt{\frac{g_{ju}^-(x_{\partial_u j})}{g_{ju}^+(x_{\partial_u j})}}$$

where we define

$$g'_{uv}(x_{\partial_u i}) \;:=\; 2\sqrt{\frac{g_{iu}^-(x_{\partial_u i})}{g_{iu}^+(x_{\partial_u i})}} \left(\frac{1}{g_{iu}^+(x_{\partial_u i})} \frac{\partial g_{iu}^+(x_{\partial_u i})}{\partial x_v} - \frac{1}{g_{iu}^-(x_{\partial_u i})} \frac{\partial g_{iu}^-(x_{\partial_u i})}{\partial x_v} \right).$$

Further, we make the bound independent of $x_{\partial_u i} \in [-1,1]^{\partial_u i}$ by taking the maximum of $|g'_{uv}(x_{\partial_u i})|$, i.e.,

$$\left|\frac{\partial h_u}{\partial x_v}(x)\right| \le \eta_i(A_i) \prod_{j \in \partial u\,:\, j \neq i} \sqrt{\frac{g_{ju}^-(x_{\partial_u j}; A_j)}{g_{ju}^+(x_{\partial_u j}; A_j)}} \tag{S17}$$

where we define

$$\eta_i(A_i) := \max_{x_{\partial_u i} \in [-1,1]^{\partial_u i}} g'_{uv}(x_{\partial_u i}; A_i)\,.$$

Now we apply the mean value theorem with (S17) to bound $|X_u - Y_u| = |h_u(X_{\partial^2 u}) - h_u(Y_{\partial^2 u})|$ by $|X_v - Y_v|$ of $v \in \partial^2 u$. It follows that for given $X_{\partial^2 u}$ and $Y_{\partial^2 u}$, there exists $\lambda' \in [0,1]$ such that

$$\begin{aligned}
&|X_u - Y_u| = |h_u(X_{\partial^2 u}) - h_u(Y_{\partial^2 u})| \\
&\le \sum_{i \in \partial u} \sum_{v \in \partial_u i} |X_v - Y_v| \left| \frac{\partial h_u}{\partial x_v} (\lambda' X_{\partial^2 u} + (1-\lambda') Y_{\partial^2 u}) \right| \\
&\le \sum_{i \in \partial u} \sum_{v \in \partial_u i} |X_v - Y_v| \eta_i(A_i) \prod_{j \in \partial u : j \neq i} \max_{\lambda \in [0,1]} \left\{ \sqrt{\frac{g_{ju}^-(\lambda X_{\partial_u j} + (1-\lambda) Y_{\partial_u j}; A_j)}{g_{ju}^+(\lambda X_{\partial_u j} + (1-\lambda) Y_{\partial_u j}; A_j)}} \right\}.
\end{aligned} \tag{S18}$$

where for the first and last inequalities, we use the mean value theorem and (S17), respectively. We note that each term in an element of the summation in the RHS of (S18) is independent to each other. Thus, it follows that

$$\begin{aligned}
&\mathbb{E}_{\tilde{\sigma}^2}\left[|X_u - Y_u|\right] \\
&\le \sum_{i \in \partial u} \sum_{v \in \partial_u i} \mathbb{E}_{\tilde{\sigma}^2}\left[|X_v - Y_v|\right] \mathbb{E}_{\tilde{\sigma}^2}\left[\eta_i(A_i)\right] \prod_{j \in \partial u : j \neq i} \mathbb{E}_{\tilde{\sigma}^2}\left[\max_{\lambda \in [0,1]} \Gamma_{ju}(\lambda X_{\partial_u j} + (1-\lambda) Y_{\partial_u j}) \right]
\end{aligned} \tag{S19}$$

where we define function $\Gamma_{iu}(x_{\partial_u i}; A_i)$ for given $x_{\partial_u i} \in [-1,1]^{\partial_u i}$ as follows:

$$\Gamma_{iu}(x_{\partial_u i}) := \sqrt{\frac{g_{iu}^-(x_{\partial_u i}; A_i)}{g_{iu}^+(x_{\partial_u i}; A_i)}}\,.$$

Note that the assumption on $\sigma^2_{\min}$ and $\sigma^2_{\max}$, i.e., $\sigma^2_{\min} + \varepsilon \le \sigma^2_{\max} < \frac{5}{2}\sigma^2_{\min}$. This implies

$$\left(-\frac{1}{\sigma^2_{\max}} + \frac{1}{\sigma^2_{\min}}\right)\frac{3}{2} - \frac{1}{\sigma^2_{\max}} \;<\; 0\,.$$

Hence, for constant ℓ and $\varepsilon > 0$, it is not hard to check that there is a finite constant η with respect to r such that

$$\max_{\tilde{\sigma}^2} \mathbb{E}_{\tilde{\sigma}^2}[\eta_i(A_i)] \;\le\; \eta \;<\; \infty \tag{S20}$$

where η may depend on only ε, $\sigma^2_{\min}$, and $\sigma^2_{\max}$.

In addition, we also obtain a bound of the last term of (S19), when r is sufficiently large, in the following lemma whose proof is presented in Section C.3.

Lemma 1. *For given $\tilde{\sigma}^2_{M_i} \in \mathcal{S}^{M_i}$ and $u \in M_i$, let $\tilde{\sigma}'^2_{M_i} \in \mathcal{S}^{M_i}$ be the set of $\tilde{\sigma}'^2_v$ such that $\tilde{\sigma}^2_u \neq \tilde{\sigma}'^2_u$ and $\tilde{\sigma}^2_v = \tilde{\sigma}'^2_v$ for all $v \in M_i \setminus \{u\}$. Then, there exists a constant $C'_{\ell,\varepsilon}$ such that for any $r \ge C'_{\ell,\varepsilon}$,*

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\max_{\lambda\in[0,1]} \Gamma_{iu}(\lambda X_{\partial_u i} + (1-\lambda) Y_{\partial_u i})\right] \;\le\; 1 - \frac{\Delta_{\min}}{2} \;<\; 1,$$

where we let $\Delta_{\min}$ be the square of the minimum Hellinger distance between the conditional densities of A_i given two different $\sigma'^2_{M_i}$ and $\sigma''^2_{M_i}$, i.e.,

$$\Delta_{\min} := \min_{\sigma^2_{M_i}, \sigma'^2_{M_i} \in \mathcal{S}^{M_i} : \sigma^2_v \neq \sigma'^2_v \;\exists v \in M_i} H^2(f_{A_i \mid \sigma'^2_{M_i}}, f_{A_i \mid \sigma''^2_{M_i}}) \;>\; 0\,.$$

Using the above lemma, we can find a sufficiently large constant $C_{\ell,\varepsilon} \ge C'_{\ell,\varepsilon}$ such that if $|\partial u| = r \ge C_{\ell,\varepsilon}$,

$$\begin{aligned}\prod_{j\in\partial u: j\neq i} \mathbb{E}_{\tilde{\sigma}^2}\left[\max_{\lambda\in[0,1]} \Gamma_{ju}(\lambda X_{\partial_u j} + (1-\lambda) Y_{\partial_u j})\right] &\le \eta\,(1-\psi_{\min})^{\frac{C_{\ell,\varepsilon}-2}{2}} \\ &\le \frac{1}{2(\ell-1)(C_{\ell,\varepsilon}-1)} \;\le\; \frac{1}{2(\ell-1)(r-1)}\end{aligned}$$

which implies (S12) with (S19) and completes the proof of Lemma 2.

C.3 Proof of Lemma 1

We first obtain a bound on X_v and Y_v for $v \in \partial_u i$. Noting that v is a non-leaf node in $G_{\rho,2k}$ and $|\partial v| = r-1$, Lemma 1 directly provides

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\mathbb{P}[\sigma^2_v \neq \tilde{\sigma}^2_v \,|\, A_{v,2k}]\right] \;=\; \mathbb{E}_{\tilde{\sigma}^2}\left[\frac{1-X_v}{2}\right] \;\le\; 4\exp\left(-\frac{\varepsilon^2}{8(8\varepsilon+1)\sigma^2_{\max}}(r-1)\right)\,.$$

Using Markov inequality for $\frac{1-X_v}{2} \ge 0$, it is easy to check that for any $\delta > 0$,

$$\mathbb{P}_{\tilde{\sigma}^2}\left[X_v < 1-\delta\right] \le \frac{8}{\delta}\exp\left(-\frac{\varepsilon^2}{8(8\varepsilon+1)\sigma^2_{\max}}(r-1)\right)\,. \tag{S21}$$

Note that

$$\begin{aligned}4\exp\left(-\frac{\varepsilon^2}{8(8\varepsilon+1)\sigma^2_{\max}}(r-1)\right) \;&\ge\; \mathbb{E}_{\tilde{\sigma}^2}\left[\mathbb{P}[\sigma^2_v \neq \tilde{\sigma}^2_v \,|\, A_v]\right] \\ &\ge\; \mathbb{E}_{\tilde{\sigma}^2}\left[\mathbb{P}[\sigma^2_v \neq \tilde{\sigma}^2_v \,|\, A_v, A_{-v}]\right] \;=\; \mathbb{E}_{\tilde{\sigma}^2}\left[\frac{1-Y_v}{2}\right]\,.\end{aligned}$$

Hence, we have the same bound in (S21) for Y_v, i.e.,

$$\mathbb{P}_{\tilde{\sigma}^2}\left[Y_v < 1-\delta\right] \;\le\; \frac{8}{\delta}\exp\left(-\frac{\varepsilon^2}{8(8\varepsilon+1)\sigma^2_{\max}}(r-1)\right)\,.$$

Using the assumption that $\sigma^2_{\min} + \varepsilon \le \sigma^2_{\max} < \frac{5}{2}\sigma^2_{\min}$, similarly to (S20), we can find finite constants η' and η'' with respect to r such that for all $x \in [0,1]^{\partial_u i}$,

$$\max_{\tilde{\sigma}'^2} \mathbb{E}_{\tilde{\sigma}'^2}\left[|\Gamma_{iu}(x)|\right] \;\le\; \eta'\,, \quad \text{and} \quad \max_{\tilde{\sigma}'^2} \mathbb{E}_{\tilde{\sigma}^2}\left[\left|\frac{\partial \Gamma_{iu}(x)}{\partial x_v}\right|\right] \;\le\; \eta''.$$

Then, it follows that for given $\delta > 0$,

$$\begin{aligned}
&\mathbb{E}_{\tilde{\sigma}^2}\left[\max_{\lambda\in[0,1]} \Gamma_{iu}(\lambda X_{\partial_u i} + (1-\lambda)Y_{\partial_u i})\right] \\
&\le (1 - \mathbb{P}_{\tilde{\sigma}^2}\left[X_v > 1-\delta \text{ and } Y_v > 1-\delta\ ,\ \forall v \in \partial_u i\right]) \max_{x\in[-1,1]^{\partial_u i}} \mathbb{E}_{\tilde{\sigma}^2}\left[\Gamma_{iu}(x)\right] + \max_{x\in[1-\delta,1]^{\partial_u i}} \mathbb{E}_{\tilde{\sigma}^2}[\Gamma_{iu}(x)] \\
&\le \left(\sum_{v\in\partial_u i} \mathbb{P}_{\tilde{\sigma}^2}[X_v \le 1-\delta] + \mathbb{P}_{\tilde{\sigma}^2}[Y_v \le 1-\delta]\right) \max_{x\in[-1,1]^{\partial_i u}} \mathbb{E}_{\tilde{\sigma}^2}[\Gamma_{iu}(x)] + \max_{x\in[1-\varepsilon,1]^{\partial_i u}} \mathbb{E}_{\tilde{\sigma}^2}[\Gamma_{iu}(x)] \qquad \text{(S22)} \\
&\le r\eta'\frac{8}{\delta}\exp\left(-\frac{\varepsilon^2}{8(8\varepsilon+1)\sigma^2_{\max}}(r-1)\right) + \max_{x\in[1-\delta,1]^{\partial_i u}} \mathbb{E}_{\tilde{\sigma}^2}[\Gamma_{iu}(x)] \qquad \text{(S23)} \\
&\le r\eta'\frac{8}{\delta}\exp\left(-\frac{\varepsilon^2}{8(8\varepsilon+1)\sigma^2_{\max}}(r-1)\right) + \delta\eta'' + \mathbb{E}_{\tilde{\sigma}^2}[\Gamma_{iu}(1_{\partial_u i})] \qquad \text{(S24)}
\end{aligned}$$

where for (S22), (S23), and (S24), we use the union bound, (S21), and the mean value theorem, respectively. We will show there exists constant Δ such that $\mathbb{E}_{\tilde{\sigma}^2}[\Gamma_{iu}(1_{\partial_u i})] \le 1-\Delta$, since the first term in (S24) is exponentially decreasing with respect to r thus there exists a constant $C'_{\ell,\varepsilon}$ such that for $r \ge C'_{\ell,\varepsilon}$,

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\max_{\lambda\in[0,1]} \Gamma_{iu}(\lambda X_{\partial_u i} + (1-\lambda)Y_{\partial_u i})\right] \ \le\ 1 - \frac{\Delta}{2}\ .$$

Recalling the property of g^+_{iu} and g^-_{iu} in (S13) and (S14), it directly follows that

$$\begin{aligned}
&\mathbb{E}_{\tilde{\sigma}^2}[\Gamma_{iu}(1_{\partial_u i})] \\
&= \int_{\mathbb{R}^{d\times M_i}} f_{A_i}[x_i \,|\, \sigma^2_{M_i} = \tilde{\sigma}^2_{M_i}] \sqrt{\frac{g^-_{iu}(1_{\partial_u i}; A_i = x_i)}{g^+_{iu}(1_{\partial_u i}; A_i = x_i)}}\, dx_i \\
&= \int_{\mathbb{R}^{d\times M_i}} f_{A_i}[x_i \,|\, \sigma^2_{M_i} = \tilde{\sigma}^2_{M_i}] \sqrt{\frac{f_{A_i}[x_i \,|\, \sigma^2_{M_i\setminus\{u\}} = \tilde{\sigma}^2_{M_i\setminus\{u\}}, \sigma^2_u = \tilde{\sigma}'^2_u]}{f_{A_i}[x_i \,|\, \sigma^2_{M_i} = \tilde{\sigma}^2_{M_i}]}}\, dx_i \\
&= \int_{\mathbb{R}^{d\times M_i}} \sqrt{f_{A_i}[x_i \,|\, \sigma^2_{M_i} = \tilde{\sigma}^2_{M_i}]} \sqrt{f_{A_i}[x_i \,|\, \sigma^2_{M_i\setminus\{u\}} = \tilde{\sigma}^2_{M_i\setminus\{u\}}, \sigma^2_u = \tilde{\sigma}'^2_u]}\, dx_i\ .
\end{aligned}$$

For notational simplicity, we define

$$\Delta(\tilde{\sigma}^2_{M_i}, \tilde{\sigma}'^2_{M_i}) := \frac{1}{2} - \frac{1}{2}\int_{\mathbb{R}^{d\times M_i}} \sqrt{f_{A_i}[x_i \,|\, \sigma^2_{M_i} = \tilde{\sigma}^2_{M_i}]} \sqrt{f_{A_i}[x_i \,|\, \sigma^2_{M_i\setminus\{u\}} = \tilde{\sigma}^2_{M_i\setminus\{u\}}, \sigma^2_u = \tilde{\sigma}'^2_u]}\, dx_i\ .$$

Then $2\Delta(\tilde{\sigma}^2_{M_i}, \tilde{\sigma}'^2_{M_i})$ is equal to the square of the Hellinger distance H between the conditional densities of A_i given $\sigma^2_{M_i} = \tilde{\sigma}^2_{M_i}$ and $\sigma^2_{M_i} = \tilde{\sigma}'^2_{M_i}$, i.e.,

$$\Delta(\tilde{\sigma}^2_{M_i}, \tilde{\sigma}'^2_{M_i}) \ =\ H^2(f_{A_i \,|\, \tilde{\sigma}^2_{M_i}}, f_{A_i \,|\, \tilde{\sigma}'^2_{M_i}}) \ >\ 0\ .$$

This implies $\Delta(\tilde{\sigma}^2_{M_i}, \tilde{\sigma}'^2_{M_i}) > 0$ and taking the minimum Δ, we complete the proof of Lemma 1.

C.4 Proof of inequality (10)

Noting that $\hat{\mu}_i^{\mathrm{BI}(k)}(A)$ is the weighted sum of $\bar{\mu}_i(A_i, \sigma'^2_{M_i})$ as described in (7), we can rewrite $\|\hat{\mu}_i^{\mathrm{BI}(k)}(A) - \mu_i\|_2^2$ as follows:

$$\|\hat{\mu}_i^{\mathrm{BI}(k)}(A) - \mu_i\|_2^2 = \sum_{\sigma'^2_{M_i}} \sum_{\sigma''^2_{M_i}} \left(\bar{\mu}_i(A_i, \sigma'^2_{M_i}) - \mu_i\right)^\top \left(\bar{\mu}_i(A_i, \sigma''^2_{M_i}) - \mu_i\right) b_i^k(\sigma'^2_{M_i}) b_i^k(\sigma''^2_{M_i})\ .$$

Hence, using Cauchy-Schwarz inequality for random variables for the summation over all $\sigma'^2_{M_i}, \sigma''^2_{M_i} \in \mathcal{S}^\ell$ except $\sigma'^2_{M_i} \neq \tilde{\sigma}^2_{M_i}$, it follows that

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\|\hat{\mu}_i^{\mathrm{BI}(k)}(A) - \mu_i\|_2^2\right] \le \mathbb{E}_{\tilde{\sigma}^2}\left[\left\|\left(\bar{\mu}_i\left(A_i, \tilde{\sigma}^2_{M_i}\right) - \mu_i\right)\right\|_2^2\right] + \sum_{\sigma''^2_{M_i}} \sum_{\sigma'^2_{M_i} \neq \tilde{\sigma}^2_{M_i}} \sqrt{\mathbb{E}_{\tilde{\sigma}^2}\left[\left(b_i^k(\sigma'^2_{M_i}) b_i^k(\sigma''^2_{M_i})\right)^2\right]}$$

$$\times \sqrt{\mathbb{E}_{\tilde{\sigma}^2}\left[\left(\left(\bar{\mu}_i(A_i,\sigma'^2_{M_i})-\mu_i\right)^\top\left(\bar{\mu}_i(A_i,\sigma''^2_{M_i})-\mu_i\right)\right)^2\right]}\,. \tag{S25}$$

Noting that the conditional density of $X=(\bar{\mu}_i(A_i,\tilde{\sigma}^2_{M_i})-\mu_i)$ given $\sigma^2=\tilde{\sigma}^2$ is identical to $\phi(X\,|\,0,\bar{\sigma}_i^2(\tilde{\sigma}^2_{M_i}))$, it follows that

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\left\|\left(\bar{\mu}_i\left(A_i,\tilde{\sigma}^2_{M_i}\right)-\mu_i\right)\right\|_2^2\right]=d\bar{\sigma}_i^2(\tilde{\sigma}^2_{M_i})\,. \tag{S26}$$

To complete the proof of (10), we hence obtain an upper bound of the last term in the RHS of (S25). For any $\sigma'^2_{M_i}\in\mathcal{S}^{M_i}$, the conditional density of the random vector $\bar{\mu}_i(A_i,\sigma'^2_{M_i})-\mu_i$ conditioned on $\sigma^2=\tilde{\sigma}^2$ is identical to

$$f_{\bar{\mu}_i(A_i,\sigma'^2_{M_i})-\mu_i}[x\,|\,\sigma^2=\tilde{\sigma}^2]=\phi\left(x\;\middle|\;0,\left(\bar{\sigma}_i^2(\sigma'^2_{M_i})\right)^2\left(\frac{1}{\tau^2}+\sum_{u\in M_i}\frac{\tilde{\sigma}_u^2}{\sigma'^4_u}\right)\right)\,.$$

Using this with some linear algebra, it is straightforward to check that for all $\sigma'^2_{M_i}\in\mathcal{S}^{M_i}$,

$$\begin{aligned}
\mathbb{E}_{\tilde{\sigma}^2}\left[\|\bar{\mu}_i(A_i,\sigma'^2_{M_i})-\mu_i\|_2^4\right]&=d(2+d)\left(\left(\bar{\sigma}_i^2(\sigma'^2_{M_i})\right)^2\left(\frac{1}{\tau^2}+\sum_{u\in M_i}\frac{\tilde{\sigma}_u^2}{\sigma'^4_u}\right)\right)^2\\
&=d(2+d)\left(\frac{\frac{1}{\tau^2}+\sum_{u\in M_i}\frac{\tilde{\sigma}_u^2}{\sigma'^4_u}}{\left(\frac{1}{\tau^2}+\sum_{u\in M_i}\frac{1}{\sigma'^2_u}\right)^2}\right)^2\\
&\le d(2+d)\left(\frac{\frac{1}{\tau^2}+\ell\frac{\sigma^2_{\max}}{\sigma^4_{\min}}}{\left(\frac{1}{\tau^2}+\ell\frac{1}{\sigma^2_{\min}}\right)^2}\right)^2
\end{aligned}$$

where for the last inequality, we use the fact that $|M_i|=\ell$ and $\sigma^2_{\min}\le\sigma_s^2\le\sigma^2_{\max}$ for any $1\le s\le S$. Using Cauchy-Schwarz inequality with the above bound, it is not hard to check that for any $\sigma'^2_{M_i},\sigma''^2_{M_i}\in\mathcal{S}^{M_i}$,

$$\begin{aligned}
&\mathbb{E}_{\tilde{\sigma}^2}\left[\left(\left(\bar{\mu}_i(A_i,\sigma'^2_{M_i})-\mu_i\right)^\top\left(\bar{\mu}_i(A_i,\sigma''^2_{M_i})-\mu_i\right)\right)^2\right]\\
\le\;&\mathbb{E}_{\tilde{\sigma}^2}\left[\left\|\bar{\mu}_i(A_i,\sigma'^2_{M_i})-\mu_i\right\|_2^2\left\|\bar{\mu}_i(A_i,\sigma''^2_{M_i})-\mu_i\right\|_2^2\right]\\
\le\;&\sqrt{\mathbb{E}_{\tilde{\sigma}^2}\left[\left\|\bar{\mu}_i(A_i,\sigma'^2_{M_i})-\mu_i\right\|_2^4\right]}\sqrt{\mathbb{E}_{\tilde{\sigma}^2}\left[\left\|\bar{\mu}_i(A_i,\sigma''^2_{M_i})-\mu_i\right\|_2^4\right]}\\
\le\;&d(2+d)\left(\frac{\frac{1}{\tau^2}+\ell\frac{\sigma^2_{\max}}{\sigma^4_{\min}}}{\left(\frac{1}{\tau^2}+\ell\frac{1}{\sigma^2_{\min}}\right)^2}\right)^2\,.
\end{aligned} \tag{S27}$$

Combining (S25), (S26) and (S27), we have

$$\mathbb{E}_{\tilde{\sigma}^2}\left[\|\hat{\mu}_i^{\mathrm{BI}(k)}(A)-\mu_i\|_2^2\right]\le d\bar{\sigma}_i^2(\tilde{\sigma}^2_{M_i})+\sqrt{d(2+d)}\left(\frac{\frac{1}{\tau^2}+\ell\frac{\sigma^2_{\max}}{\sigma^4_{\min}}}{\left(\frac{1}{\tau^2}+\ell\frac{1}{\sigma^2_{\min}}\right)^2}\right)\sum_{\sigma''^2_{M_i}}\sum_{\sigma'^2_{M_i}\neq\tilde{\sigma}^2_{M_i}}\sqrt{\mathbb{E}_{\tilde{\sigma}^2}\left[\left(b_i^k(\sigma'^2_{M_i})b_i^k(\sigma''^2_{M_i})\right)^2\right]}\,. \tag{S28}$$

Using Cauchy-Schwarz inequality and Jensen's inequality sequentially, it follows that

$$\begin{aligned}
\sum_{\sigma''^2_{M_i}}\sum_{\sigma'^2_{M_i}\neq\tilde{\sigma}^2_{M_i}}\left(\mathbb{E}_{\tilde{\sigma}^2}\left[\left(b_i^k(\sigma'^2_{M_i})b_i^k(\sigma''^2_{M_i})\right)^2\right]\right)^{1/2}&\le\sum_{\sigma''^2_{M_i}}\sum_{\sigma'^2_{M_i}\neq\tilde{\sigma}^2_{M_i}}\left(\mathbb{E}_{\tilde{\sigma}^2}\left[\left(b_i^k(\sigma'^2_{M_i})\right)^4\right]\right)^{1/4}\left(\mathbb{E}_{\tilde{\sigma}^2}\left[\left(b_i^k(\sigma''^2_{M_i})\right)^4\right]\right)^{1/4}\\
&=\left(\sum_{\sigma'^2_{M_i}\neq\tilde{\sigma}^2_{M_i}}\left(\mathbb{E}_{\tilde{\sigma}^2}\left[\left(b_i^k(\sigma'^2_{M_i})\right)^4\right]\right)^{1/4}\right)\left(\sum_{\sigma''^2_{M_i}}\left(\mathbb{E}_{\tilde{\sigma}^2}\left[\left(b_i^k(\sigma''^2_{M_i})\right)^4\right]\right)^{1/4}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \left(\sum_{\sigma'^2_{M_i} \neq \tilde{\sigma}^2_{M_i}} \mathbb{E}_{\tilde{\sigma}^2} \left[\left(b_i^k(\sigma'^2_{M_i}) \right)^4 \right] \right)^{1/4} \left(\sum_{\sigma''^2_{M_i}} \mathbb{E}_{\tilde{\sigma}^2} \left[\left(b_i^k(\sigma''^2_{M_i}) \right)^4 \right] \right)^{1/4} \\
&\leq \left(\sum_{\sigma'^2_{M_i} \neq \tilde{\sigma}^2_{M_i}} \mathbb{E}_{\tilde{\sigma}^2} \left[b_i^k(\sigma'^2_{M_i}) \right] \right)^{1/4} \left(\sum_{\sigma''^2_{M_i}} \mathbb{E}_{\tilde{\sigma}^2} \left[b_i^k(\sigma''^2_{M_i}) \right] \right)^{1/4} \\
&= \left(1 - \mathbb{E}_{\tilde{\sigma}^2} \left[b_i^k(\tilde{\sigma}^2_{M_i}) \right] \right)^{1/4} ,
\end{aligned}
$$

where for the last inequality and the last equality, we use the fact that b_i^k is normalized, i.e., $0 \leq b_i^k(\sigma^2_{M_i}) \leq 1$ and $\sum_{\sigma^2_{M_i}} b_i^k(\sigma^2_{M_i}) = 1$. This completes the proof of (10) with (S28).

C.5 Proof of Inequality (15)

We start with rewriting the difference between MSE's of $\hat{\mu}_\tau^{\mathsf{ora}(k)}(A)$ and $\hat{\mu}_\tau^{\mathrm{BI}(k)}(A)$ for $\tau \in V$ as follows:

$$
\begin{aligned}
&\|\hat{\mu}_\tau^{\mathsf{ora}(k)}(A) - \mu_\tau\|_2^2 - \|\hat{\mu}_\tau^{\mathrm{BI}(k)}(A) - \mu_\tau\|_2^2 \\
&= \sum_{\sigma'^2_{M_\tau}, \sigma''^2_{M_\tau} \in \mathcal{S}^\ell} \left(\mathbb{P}[\sigma^2_{M_\tau} = \sigma'^2_{M_\tau} \mid A, \sigma^2_{\partial W_{\tau,2k+1}}] \mathbb{P}[\sigma^2_{M_\tau} = \sigma''^2_{M_\tau} \mid A, \sigma^2_{\partial W_{\tau,2k+1}}] - b_\tau^k(\sigma'^2_{M_\tau}) b_\tau^k(\sigma''^2_{M_\tau}) \right) \\
&\quad \times \left(\bar{\mu}_\tau \left(A_\tau, \sigma'^2_{M_\tau} \right) - \mu_\tau \right)^\top \left(\bar{\mu}_\tau \left(A_\tau, \sigma''^2_{M_\tau} \right) - \mu_\tau \right) .
\end{aligned}
$$

Then, using Cauchy-Schwarz inequality for random variables X and Y, i.e., $|\,\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(\mathrm{MSE}(\hat{\mu}_\tau^{\mathsf{ora}(k)}(A)) - \mathrm{MSE}(\hat{\mu}_\tau^{\mathrm{BI}(k)}(A)) \right) \right] \\
&\leq \sum_{\sigma^2_{M_\tau}, \sigma'^2_{M_\tau} \in \mathcal{S}^\ell} \sqrt{ \mathbb{E}\left[\left(\mathbb{P}[\sigma^2_{M_\tau} = \sigma'^2_{M_\tau} \mid A, \sigma^2_{\partial W_{\tau,2k+1}}] \mathbb{P}[\sigma^2_{M_\tau} = \sigma''^2_{M_\tau} \mid A, \sigma^2_{\partial W_{\tau,2k+1}}] - b_\tau^k(\sigma'^2_{M_\tau}) b_\tau^k(\sigma''^2_{M_\tau}) \right)^2 \right] } \\
&\quad \times \sqrt{ \mathbb{E}\left[\left(\left(\bar{\mu}_\tau(A_\tau, \sigma^2_{M_\tau}) - \mu_i \right)^\top \left(\bar{\mu}_\tau(A_\tau, \sigma'^2_{M_\tau}) - \mu_i \right) \right)^2 \right] }
\end{aligned}
$$

which completes the proof of (15) with (S27).

References

M. Everingham, SM. A. Eslami, L. Van Gool, C. KI Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

R. Girshick. Fast r-cnn. In *Proc. of the IEEE ICCV*, 2015.

H. Han, C. Otto, X. Liu, and A. K. Jain. Demographic estimation from face images: Human vs. machine performance. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1148–1161, 2015.

A. Lanitis. Comparative evaluation of automatic age-progression methodologies. *EURASIP Journal on Advances in Signal Processing*, 2008:101, 2008.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *Proc. of ECCV*, 2016.

G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes. Overview of research on facial ageing using the FG-NET ageing database. *IET Biometrics*, 5(2):37–46, 2016.

S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. of NIPS*, 2015.

R. Rothe, R. Timofte, and L. Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proc. of ICCV*, pages 10–15, 2015.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.