SUPPLEMENT TO "EXACT EXPONENT IN OPTIMAL RATES FOR CROWDSOURCING"

Chao Gao, Yu Lu and Dengyong Zhou

A Proof of Corollary 3.1

Proof. Under the assumption that $mI(\pi)\to\infty$, the upper bound is a special case of Theorem 3.1. Note that

$$\begin{aligned} & I(\pi) \\ = & -\min_{0\leq t\leq 1}\frac{1}{m}\sum_{i=1}^{m}\log\left(p_i^{1-t}(1-p_i)^t+p_i^{1-t}(1-p_i)^t\right) \\ = & -\frac{1}{m}\sum_{i=1}^{m}\log\left(2\sqrt{p_i(1-p_i)}\right) \\ = & I(p). \end{aligned}$$

We focus on the proof of the lower bound, which involves weaker assumptions than that of Theorem 3.1. Using a similar analysis as (15)-(16), we have

$$\begin{aligned} & \inf_{\hat{y}}\sup_{y\in\{1,2\}^n}\mathbb{E}L(\hat{y},y) \\ \geq & \frac{1}{n}\sum_{j=1}^{n}\inf_{\hat{y}_j}\left[\frac{1}{2}\mathbb{P}_1\{\hat{y}_j=2\}+\frac{1}{2}\mathbb{P}_2\{\hat{y}_j=1\}\right]. \end{aligned}$$

Following the proof of Theorem 3.1 with the confusion matrix $\pi^{(i)}$ replaced by (5), we have

$$\begin{aligned} & \inf_{\hat{y}_j}\left[\frac{1}{2}\mathbb{P}_1\{\hat{y}_j=2\}+\frac{1}{2}\mathbb{P}_2\{\hat{y}_j=1\}\right] \\ \geq & \exp\left(-mI(p)\right)e^{-L}\mathbb{Q}(0<S_m<L), \end{aligned}$$

where $S_m=\sum_{i\in[m]}W_i$, and under the distribution $\mathbb{Q}$,

$$\mathbb{Q}_i\left(W_i=\frac{1}{2}\log\frac{1-p_i}{p_i}\right)=\mathbb{Q}_i\left(W_i=\frac{1}{2}\log\frac{p_i}{1-p_i}\right)=\frac{1}{2}.$$

Therefore, S_m has a symmetric distribution around 0. Letting $L=2\sqrt{\mathrm{Var}_Q(S_m)}$, we have

$$\mathbb{Q}(0<S_m<L)\geq\frac{1}{2}-\mathbb{Q}(S_m\geq L)\geq\frac{1}{2}-\frac{\mathrm{Var}_Q(S_m)}{L^2}\geq\frac{1}{4}.$$

Finally, we need to show that $L=o(mI(p))$. We claim that

$$\begin{aligned} & \sum_{i=1}^{m}\mathrm{Var}_QW_i=\frac{1}{4}\sum_{i=1}^{m}\left(\log\frac{1-p_i}{p_i}\right)^2 \\ \leq & -8\max_{1\leq i\leq m}\left(|\log(p_i)|\vee|\log(1-p_i)|\vee 2\right) \\ & \sum_{i=1}^{m}\log\left(2\sqrt{p_i(1-p_i)}\right). \end{aligned}$$

This is becase when $p_i \in [1/16, 15/16]$, we have $\left|\log \frac{1-p_i}{p_i}\right|^2 \leq 6(2p_i-1)^2 \leq -6\log(4p_i(1-p_i))$. When $p_i \in (0, 1/16) \cup (15/16, 1)$, $\left|\log \frac{1-p_i}{p_i}\right| \leq -2\log(4p_i(1-p_i))$ and $\left|\log \frac{1-p_i}{p_i}\right| \leq 2|\log(p_i)| \vee 2|\log(1-p_i)|$. Therefore, under the assumption that

$$\max_{1 \leq i \leq m} (|\log(p_i)| \vee |\log(1-p_i)|) = o(mI(p)),$$

$L = o(mI(p))$ holds, and the proof is complete. □

B Proof of Lemma 6.1

Let $f(t) = \sum_{i=1}^m \log B_t(\pi_{1*}^{(i)}, \pi_{2*}^{(i)})$. Then we have $f'(t_0) = 0$ by its definition. First, we are going to prove $0 < t_0 < 1$. The concavity of logarithm gives us $x^t y^{1-t} \leq tx + (1-t)y$ for non-negative x, y and $t \in [0, 1]$, which implies

$$f(t) = \sum_{i \in [m]} \log B_t(\pi_{1*}^{(i)}, \pi_{2*}^{(i)}) \leq \sum_{i \in [m]} \log \left(\sum_{h=1}^k \left((1-t)\pi_{1h}^{(i)} + t\pi_{2h}^{(i)} \right) \right) = 0.$$

For $t \in (0, 1)$, the equality holds if and only if $\pi_{1h}^{(i)} = \pi_{2h}^{(i)}$ for all $h \in [k]$ and $i \in [m]$. As there is at least one non-spammer, we must have $f(t) < 0 = f(0) = f(1)$ for $t \in (0, 1)$. Hence the minimizer $t_0 \in (0.1)$.

Now we are going to show the uniqueness of t_0 by proving that

$$f''(t) = \mathrm{Var}(S_m) > 0, \ \forall t \in (0, 1)$$

where $S_m = \sum_{i \in [m]} W_i$. To simplify the notation, let us define $w_{ih} = t \log \left(\frac{\pi_{2h}^{(i)}}{\pi_{1h}^{(i)}} \right)$ and $p_{ih} = \left(\pi_{1h}^{(i)}\right)^{1-t} \left(\pi_{2h}^{(i)}\right)^t$ for all $i \in [m]$ and $h \in [k]$. Now $\mathbb{Q}_i(W_i = w_{ih}) = p_{ih} / \sum_h p_{ih}$ and $B_t(\pi_{1*}^{(i)}, \pi_{2*}^{(i)}) = \sum_{h \in [k]} p_{ih}$. Notice that $\frac{d}{dt} p_{ih} = p_{ih} w_{ih}$, we have

$$\frac{d}{dt} f(t) = \sum_{i \in [m]} \frac{\sum_h p_{ih} w_{ih}}{\sum_h p_{ih}} = \sum_{i \in [m]} \mathbb{E} W_i = \mathbb{E} S_m, \tag{19}$$

and

$$\frac{d^2}{dt^2} f(t) = \sum_{i \in [m]} \frac{\sum_h p_{ih} w_{ih}^2 \sum_h p_{ih} - \left(\sum_h p_{ih} w_{ih}\right)^2}{\left(\sum_h p_{ih}\right)^2} = \sum_{i \in [m]} \mathrm{Var}(W_i) = \mathrm{Var}(S_m). \tag{20}$$

Since the set $\mathcal{A}_\alpha$ is non-empty, there is at least one $\mathrm{Var}(W_i) > 0$. Thus, $f''(t) = \mathrm{Var}(S_m) > 0$.

C Proof of Lemma 6.2

From (19), we know $\mathbb{E} S_m = f'(t_0) = 0$. Since $t_0 > 0$ by lemma 6.1, we can rescale W_i by $W_i / (-t_0 \log \rho_m)$ and the value of $S_m / \sqrt{\mathrm{Var}(S_m)}$ will not change. Let us define $V_i =$

$W_i/(-t_0 \log \rho_m)$ and $R_m = \sum_{i=1}^m V_i$. Then we have $|V_i| \leq 1$. To prove a central limit theorem of S_m, it is sufficient to check the following Lindeberg's condition [**?**], that is, for any $\epsilon > 0$,

$$\frac{1}{\mathrm{Var}(R_m)} \sum_{i=1}^m \mathbb{E}\left((V_i - \mathbb{E}V_i)^2 \mathbf{I}\{(V_i - \mathbb{E}V_i)^2 \geq \epsilon^2 \mathrm{Var}(R_m)\}\right) \to 0 \text{ as } m \to \infty. \tag{21}$$

Note that for a discrete random variable X who takes value x_a with probability p_a for $a \in [N]$,

$$\mathrm{Var}(X) = \left(\sum_a p_a\right)\left(\sum_a p_a x_a^2\right) - \left(\sum_a p_a x_a\right)^2 = \sum_{a,b} p_a p_b (x_a - x_b)^2.$$

Then, for any $i \in \mathcal{A}_\alpha$, we have

$$\begin{aligned}
\mathrm{Var}(V_i) &= \frac{1}{\log^2 \rho_m} \sum_{a,b} \frac{\left(\pi_{1a}^{(i)}\pi_{1b}^{(i)}\right)^{1-t}\left(\pi_{2a}^{(i)}\pi_{2b}^{(i)}\right)^{t}}{B_t^2(\pi_{1*}^{(i)}, \pi_{2*}^{(i)})} \log^2\left(\frac{\pi_{2a}^{(i)}\pi_{1b}^{(i)}}{\pi_{1a}^{(i)}\pi_{2b}^{(i)}}\right) \\
&\geq \frac{1}{\log^2 \rho_m}\left(\pi_{12}^{(i)}\pi_{11}^{(i)}\right)^{1-t}\left(\pi_{22}^{(i)}\pi_{21}^{(i)}\right)^{t} \log^2\left(\frac{\pi_{22}^{(i)}\pi_{11}^{(i)}}{\pi_{12}^{(i)}\pi_{21}^{(i)}}\right) \\
&\geq \frac{1}{\log^2 \rho_m}\left(\pi_{12}^{(i)}\pi_{11}^{(i)}\right)^{1-t}\left(\pi_{22}^{(i)}\pi_{21}^{(i)}\right)^{t} \log^2\left((1+\alpha)^2\right) \\
&\geq \frac{\rho_m^2}{\log^2 \rho_m} 4\log^2(1+\alpha) \\
&\geq \frac{\rho_m^2}{\log^2 \rho_m} \min\{\alpha^2, 1\}.
\end{aligned}$$

Here the second inequality is due to the assumption that for any $i \in \mathcal{A}$, $\pi_{aa}^{(i)} \geq \pi_{ab}^{(i)}(1+\alpha)$ for any $b \neq a$. We have used the assumption that $\pi_{ab}^{(i)} \geq \rho_m$ for the third inequality. The last inequality is because $\log(1+\alpha) \geq \alpha/(1+\alpha) \geq \min\{\alpha/2, 1/2\}$ for positive α. Take a sum of $\mathrm{Var}(V_i)$ over $i \in \mathcal{A}_\alpha$,

$$\mathrm{Var}(R_m) = \sum_{i \in [m]} \mathrm{Var}(V_i) \geq |\mathcal{A}_\alpha| \frac{\rho_m^2}{\log^2 \rho_m} \min\{\alpha^2, 1\} \geq cm \frac{\rho_m^2}{\log^2 \rho_m} \min\{\alpha^2, 1\},$$

for some constant $c \in (0,1)$. Since $(V_i - \mathbb{E}V_i)^2 \leq 2V_i^2 + 2(\mathbb{E}V_i)^2 \leq 4$, we will have

$$\mathbf{I}\{(V_i - \mathbb{E}V_i)^2 \geq \epsilon^2 \mathrm{Var}(R_m)\} = 0$$

when $4\log^2 \rho_m < \epsilon^2 |\mathcal{A}_\alpha| \rho_m^2 \min\{\alpha^2, 1\}$. Notice that $\mathbb{E}(V_i - \mathbb{E}V_i)^2 \leq 4$, we apply the Dominated Convergence Theorem to conclude

$$\mathbb{E}\left((V_i - \mathbb{E}V_i)^2 \mathbf{I}\{(V_i - \mathbb{E}V_i)^2 \geq \epsilon^2 \mathrm{Var}(R_m)\}\right) \to 0.$$

Thus, the Lindeberg condition holds when $|\log \rho_m| = o(\rho_m |\mathcal{A}_{0.01}|^{1/2})$.

D Proof of Lemma 6.3

We are first going to show $\lambda_0 \leq -2\log\rho_m$. Recall that

$$f(\lambda) = \prod_{i=1}^{m}\left((1-p_i)e^{\lambda/2} + p_i e^{-\lambda/2}\right).$$

For all $\lambda \geq -2\log\rho_m$, we have $f(\lambda) > \rho_m^{-m}\prod_{i=1}^{m}(1-p_i) \geq 1$, and $f(0) = 1$. Thus, the minimizer of $f(\lambda)$ must be in the interval $(0, -2\log\rho_m]$.

Again, we are going to prove the central limit theorem of S_m by checking the following Lindeberg's condition. For any $\epsilon > 0$,

$$\lim_{m\to\infty}\frac{1}{\mathrm{Var}_{\mathbb{Q}}(S_m)}\sum_{i=1}^{m}\mathbb{E}_{\mathbb{Q}}\left[(W_i - \mathbb{E}_{\mathbb{Q}}W_i)^2\mathbf{I}\left\{|W_i - \mathbb{E}_{\mathbb{Q}}W_i| > \epsilon\sqrt{\mathrm{Var}_{\mathbb{Q}}(S_m)}\right\}\right] = 0 \tag{22}$$

When $\lambda_0 \in (0, -2\log\rho_m]$, a lower bound of $\mathrm{Var}_Q(W_i)$ is given by

$$\mathrm{Var}_{\mathbb{Q}}(W_i) \quad = \quad \lambda_0^2\frac{p_i(1-p_i)}{\left((1-p_i)e^{\lambda_0/2} + p_i e^{-\lambda_0/2}\right)^2} \geq \lambda_0^2 e^{-\lambda}p_i(1-p_i) \geq \lambda_0^2\rho_m^2 p_i(1-p_i)$$

Therefore, $\mathrm{Var}_{\mathbb{Q}}(S_m) = \sum_{i=1}^{m}\mathrm{Var}_{\mathbb{Q}}(W_i) \geq \lambda_0^2\rho_m^2\sum_{i\in[m]}p_i(1-p_i)$. Notice that $|W_i - \mathbb{E}_{\mathbb{Q}}W_i| \leq |W_i| + \mathbb{E}|W_i| = \lambda$, for any fixed $\epsilon > 0$, we will have

$$\mathbf{I}\left\{|W_i - \mathbb{E}_{\mathbb{Q}}W_i| > \epsilon\sqrt{\mathrm{Var}_{\mathbb{Q}}(S_m)}\right\} = 0$$

when $\rho_m^2\sum_{i\in[m]}p_i(1-p_i) \to \infty$ as $m \to \infty$. Since $\mathrm{Var}_{\mathbb{Q}}(W_i)/\lambda_0^2 \leq 1/4$, the Dominated Convergence Theorem implies the desired Lindeberg's condition (22).