Appendix

1. Proof of Corollary 2.2

Recall that

$$\begin{aligned} I(a,b) &= \Pr(\theta \geq 0.5 | \theta \sim \text{Beta}(a,b)) \quad (1) \\ &= \frac{1}{B(a,b)} \int_{0.5}^{1} t^{a-1}(1-t)^{b-1} \mathrm{d}t, \end{aligned}$$

where $B(a,b)$ is the beta function.

It is easy to see that $I(a,b) > 0.5 \iff I(a,b) > 1 - I(a,b)$. We re-write $1 - I(a,b)$ as follows

$$\begin{aligned} 1 - I(a,b) &= \frac{1}{B(a,b)} \int_{0}^{0.5} t^{a-1}(1-t)^{b-1} \mathrm{d}t \\ &= \frac{1}{B(a,b)} \int_{0.5}^{1} t^{b-1}(1-t)^{a-1} \mathrm{d}t, \end{aligned}$$

where the second equality is obtained by setting $t := 1 - t$. Then we have:

$$\begin{aligned} & I(a,b) - (1 - I(a,b)) \\ &= \frac{1}{B(a,b)} \int_{0.5}^{1} (t^{a-1}(1-t)^{b-1} - t^{b-1}(1-t)^{a-1}) \mathrm{d}t \\ &= \frac{1}{B(a,b)} \int_{0.5}^{1} t^{a-1}(1-t)^{b-1} \left(\left(\frac{t}{1-t} \right)^{a-b} - 1 \right) \mathrm{d}t \end{aligned}$$

Since $t > 0.5$, $\frac{t}{1-t} > 1$. When $a > b$, $\left(\frac{t}{1-t}\right)^{a-b} > 1$ and hence $I(a,b) - (1 - I(a,b)) > 0$, i.e, $I(a,b) > 0.5$. When $a = b$, $\left(\frac{t}{1-t}\right)^{a-b} \equiv 1$ and $I(a,b) = 0.5$. When $a < b$, $\left(\frac{t}{1-t}\right)^{a-b} < 1$ and $I(a,b) < 0.5$.

2. Proof of Proposition 2.3

We use the proof technique in (Xie & Frazier, 2012) to prove Proposition 2.3. By Proposition 2.1, we first obtain the value function:

$$\begin{aligned} V(S^0) &\doteq \sup_{\pi} \mathbb{E}^{\pi} \left(\sum_{i \in H_T} \mathbf{1}(i \in H^*) + \sum_{i \notin H_T} \mathbf{1}(i \notin H^*) \Big| \mathcal{F}_T \right) \quad (2) \\ &= \sup_{\pi} \mathbb{E}^{\pi} \left(\sum_{i=1}^{K} h(P_i^T) \right). \end{aligned}$$

To decompose the final accuracy $\sum_{i=1}^{K} h(P_i^T)$ into to the intermediate reward at each stage, we define $G_0 = \sum_{i=1}^{K} h(P_i^0)$ and $G_{t+1} = \sum_{i=1}^{K} h(P_i^{t+1}) - \sum_{i=1}^{K} h(P_i^t)$. Then, $\sum_{i=1}^{K} h(P_i^T)$ can be decomposed as: $\sum_{i=1}^{K} h(P_i^T) \equiv G_0 + \sum_{t=0}^{T-1} G_{t+1}$. The value function can now be re-written as follows:

$$\begin{aligned} V(S^0) &= G_0(S^0) + \sup_{\pi} \sum_{t=0}^{T-1} \mathbb{E}^{\pi}(G_{t+1}) \\ &= G_0(S^0) + \sup_{\pi} \sum_{t=0}^{T-1} \mathbb{E}^{\pi} \left(\mathbb{E}(G_{t+1} | \mathcal{F}_t) \right). \\ &= G_0(S^0) + \sup_{\pi} \sum_{t=0}^{T-1} \mathbb{E}^{\pi} \left(\mathbb{E}(G_{t+1} | S^t, i_t) \right). \end{aligned}$$

Here, the first inequality is true because G_0 is determinant and independent of π, the second inequality is due to the tower property of conditional expectation and the third inequality holds because P_i^{t+1} and P_i^i, and thus, G_{t+1} depend on $\mathcal{F}_t$ only through S^t and i_t. We define intermediate expected reward gained by labeling the i_t-th instance at the state S^t as follows:

$$\begin{aligned} R(S^t, i_t) &= \mathbb{E}(G_{t+1} | S^t, i_t) \\ &= \mathbb{E} \left(\sum_{i=1}^{K} h(P_i^{t+1}) - \sum_{i=1}^{K} h(P_i^t) | S^t, i_t \right) \\ &= \mathbb{E} \left(h(P_{i_t}^{t+1}) - h(P_{i_t}^t) | S^t, i_t \right). \quad (3) \end{aligned}$$

The last equation is due to the fact that only $P_{i_t}^t$ will be changed if the i_t-th instance is labeled next. With the expected reward function in place, our value function takes the following form:

$$V(S^0) = G_0(\mathbf{s}) + \sup_{\pi} \mathbb{E}^{\pi} \left(\sum_{t=0}^{T-1} R(S^t, i_t) \Big| S^0 \right). \quad (4)$$

3. Proof of Proposition 3.1

To prove the failure of deterministic KG, we first show a key property for the expected reward function:

$$\begin{aligned} R(a,b) = & \frac{a}{a+b} \left(h(I(a+1,b)) - h(I(a,b)) \right) \quad (5) \\ & + \frac{b}{a+b} \left(h(I(a,b+1)) - h(I(a,b)) \right). \end{aligned}$$

Lemma 3.1. *When a, b are positive integers, if $a = b$, $R(a,b) = \frac{0.5^{2a}}{aB(a,a)}$ and if $a \neq b$, $R(a,b) = 0$.*

To prove lemma 3.1, we first present several basic properties for $B(a,b)$ and $I(a,b)$, which will be used in all the following theorems and proofs.

1. Properties for $B(a,b)$:

$$B(a,b) = B(b,a) \tag{6}$$
$$B(a+1,b) = \frac{a}{a+b}B(a,b) \tag{7}$$
$$B(a,b+1) = \frac{b}{a+b}B(a,b) \tag{8}$$

2. Properties for $B(a,b)$:

$$I(a,b) = 1 - I(b,a) \tag{9}$$
$$I(a+1,b) = I(a,b) + \frac{0.5^{a+b}}{aB(a,b)} \tag{10}$$
$$I(a,b+1) = I(a,b) - \frac{0.5^{a+b}}{bB(a,b)} \tag{11}$$

The properties for $I(a,b)$ are derived from the basic property of regularized incomplete beta function [1].

Proof. When $a = b$, by Corollary 2.2, we have $I(a+1,b) > 0.5$, $I(a,b) = 0.5$ and $I(a,b+1) < 0.5$. Therefore, the expected reward (5) takes the following form:

$$\begin{aligned} R(a,b) &= 0.5(I(a+1,a) - I(a,a)) + \\ &\quad 0.5((1 - I(a,a+1)) - I(a,a)) \\ &= I(a+1,a) - I(a,a) \\ &= \frac{0.5^{2a}}{aB(a,a)} \end{aligned}$$

When $a > b$, since a, b are integers, we have $a \geq b+1$ and hence $I(a+1,b) > 0.5, I(a,b) > 0.5, I(a,b+1) \geq 0.5$ according to Corollary 2.2. The expected reward (5) now becomes:

$$\begin{aligned} R(a,b) =& \frac{a}{a+b}I(a+1,b) + \frac{b}{a+b}I(a,b+1) - I(a,b) \\ =& \frac{a}{a+b}\frac{1}{B(a+1,b)}\int_{0.5}^{1} t \cdot t^{a-1}(1-t)^{b-1}\mathrm{d}t \\ &+ \frac{b}{a+b}\frac{1}{B(a,b+1)}\int_{0.5}^{1} t^{a-1}(1-t)(1-t)^{b-1}\mathrm{d}t \\ &- I(a,b) \\ =& \frac{1}{B(a,b)}\int_{0.5}^{1}(t+(1-t))\cdot t^{a-1}(1-t)^{b-1}\mathrm{d}t - I(a,b) \\ =& I(a,b) - I(a,b) = 0. \end{aligned}$$

[1] http://dlmf.nist.gov/8.17

Here we use (7) and (8) to show that $\frac{a}{a+b}\frac{1}{B(a+1,b)} = \frac{b}{a+b}\frac{1}{B(a,b+1)} = \frac{1}{B(a,b)}$.

When $a \leq b-1$, we can prove $R(a,b) = 0$ in a similar way. □

With Lemma 3.1 in place, the proof for Proposition 3.1 is straightforward. Recall that the deterministic KG policy chooses the next instance according to

$$i_t = \arg\max_i R(S^t, i) \equiv \arg\max_i R(a_i^t, b_i^t),$$

and breaks the tie by selecting the one with the smallest index. Since $R(a,b) > 0$ if and only if $a = b$, at the initial stage $t = 0$, $R(a_i^0, b_i^0) > 0$ for those instances $i \in \mathcal{E} = \{i : a_i^0 = b_i^0\}$. The policy will first select $i_0 \in \mathcal{E}$ with the largest $R(a_i^0, b_i^0)$. After obtaining the label y_{i_0}, either $a_{i_0}^0$ or $b_{i_0}^0$ will add one and hence $a_{i_0}^1 \neq b_{i_0}^1$ and $R(a_{i_0}^1, b_{i_0}^1) = 0$. The policy will select another instance $i_1 \in \mathcal{E}$ with the "current" largest expected reward and the expected reward for i_1 after obtaining the label y_{i_1} will then become zero. As a consequence, the KG policy will label each instance in $\mathcal{E}$ for the first $|\mathcal{E}|$ stages and $R(a_i^{|\mathcal{E}|}, b_i^{|\mathcal{E}|}) = 0$ for all $i \in \{1, \ldots, K\}$. Then the deterministic policy will break the tie selecting the first instance to label. From now on, for any $t \geq |\mathcal{E}|$, if $a_1^t \neq b_1^t$, then the expected reward $R(a_1^t, b_1^t) = 0$. Since the expected reward for other instances are all zero, the policy will still label the first instance. On the other hand, if $a_1^t = b_1^t$, and the first instance is the only one with the positive expected reward and the policy will label it. Thus Proposition 3.1 is proved.

Remark. *For randomized KG, after getting one label for each instance in $\mathcal{E}$ for the first $|\mathcal{E}|$ stages, the expected reward for each instance has become zero. Then randomized KG will uniformly select one instance to label. At any stage $t \geq |\mathcal{E}|$, if there exists one instance i (at most one instance) with $a_i^t = b_i^t$, the KG policy will provide the next label for i; otherwise, it will randomly select an instance to label.*

4. Proof of Theorem 3.2

To prove the consistency of the optimistic KG policy, we first show the exact values for $R_\alpha^+(a,b) = \max(R_1(a,b), R_2(a,b))$.

1. When $a \geq b+1$:

$$R_1(a,b) = I(a+1,b) - I(a,b) = \frac{0.5^{a+b}}{aB(a,b)} > 0;$$

$$R_2(a,b) = I(a,b+1) - I(a,b) = -\frac{0.5^{a+b}}{bB(a,b)} < 0.$$

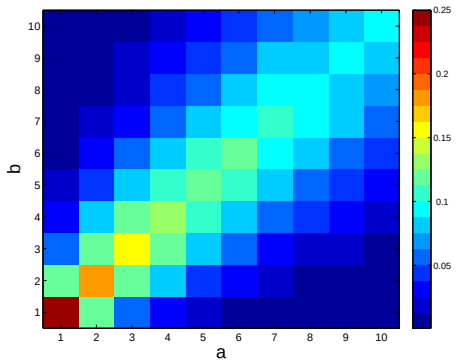

Figure 1. Illustration of $R^+(a,b)$.

Therefore,

$$R^+(a,b) = R_1(a,b) = \frac{0.5^{a+b}}{aB(a,b)} > 0.$$

2. When $a = b$:

$$R_1(a,b) = I(a+1,a) - I(a,a) = \frac{0.5^{2a}}{aB(a,a)};$$
$$R_2(a,b) = 1 - I(a,a+1) - I(a,a) = \frac{0.5^{2a}}{aB(a,a)}.$$

Therefore, we have $R_1 = R_2$ and

$$R^+(a,b) = R_1(a,b) = R_2(a,b) = \frac{0.5^{2a}}{aB(a,a)} > 0.$$

3. When $b - 1 \geq a$:

$$R_1(a,b) = I(a,b) - I(a+1,b) = -\frac{0.5^{a+b}}{aB(a,b)} < 0;$$
$$R_2(a,b) = I(a,b) - I(a,b+1) = \frac{0.5^{a+b}}{bB(a,b)} > 0.$$

Therefore

$$R^+(a,b) = R_2(a,b) = \frac{0.5^{a+b}}{bB(a,b)} > 0.$$

For better visualization, we plot values of $R^+(a,b)$ for different a, b in Figure 1.

As we can see $R^+(a,b) > 0$ for any positive integers (a,b), we first prove that $\lim_{a+b\to\infty} R^+(a,b) = 0$ in the following Lemma.

Lemma 4.1. *Properties for $R^+(a,b)$:*

1. *$R(a,b)$ is symmetric, i.e., $R^+(a,b) = R^+(b,a)$.*
2. *$\lim_{a\to\infty} R^+(a,a) = 0$.*
3. *For any fixed $a \geq 1$, $R^+(a+k, a-k) = R^+(a-k, a+k)$ is monotonically decreasing in k for $k = 0, \ldots, a-1$.*
4. *When $a \geq b$, for any fixed b, $R^+(a,b)$ is monotonically decreasing in a. By the symmetry of $R^+(a,b)$, when $b \geq a$, for any fixed a, $R^+(a,b)$ is monotonically decreasing in b.*

By the above four properties, we have $\lim_{(a+b)\to\infty} R^+(a,b) = 0$.

Proof. We first prove these four properties.

- Property 1: By the fact that $B(a,b) = B(b,a)$, the symmetry of $R^+(a,b)$ is straightforward.
- Property 2: For $a > 1$, $\frac{R^+(a,a)}{R^+(a-1,a-1)} = \frac{2a-1}{2a} < 1$ and hence $R^+(a,a)$ is monotonically decreasing in a. Moreover,

$$\begin{aligned} R^+(a,a) &= R^+(1,1)\prod_{i=2}^{a}\frac{2i-1}{2i} \\ &= R^+(1,1)\prod_{i=2}^{a}(1-\frac{1}{2i}) \\ &\leq R^+(1,1)e^{-\sum_{i=2}^{a}\frac{1}{2i}} \end{aligned}$$

  Since $\lim_{a\to\infty}\sum_{i=2}^{a}\frac{1}{2i} = \infty$ and $R^+(a,a) \geq 0$, $\lim_{a\to\infty} R^+(a,a) = 0$.

- Property 3: For any $k \geq 0$,

$$\begin{aligned} &\frac{R^+(a+(k+1), a-(k+1))}{R^+(a+k, a-k)} \\ =&\frac{(a+k)B(a+k, a-k)}{(a+k+1)B(a+(k+1), a-(k+1))} \\ =&\frac{a-(k+1)}{a+(k+1)} < 1. \end{aligned}$$

- Property 4: When $a \geq b$, for any fixed b:

$$\begin{aligned} \frac{R^+(a+1,b)}{R^+(a,b)} =&\frac{aB(a,b)}{2(a+1)B(a+1,b)} \\ =&\frac{a(a+b)}{2a(a+1)} < 1. \end{aligned}$$

According to the third property, when $a+b$ is an even number, we have $R^+(a,b) < R^+(\frac{a+b}{2}, \frac{a+b}{2})$. According to the fourth property, when $a+b$ is an odd number and $a \geq b+1$, we have $R^+(a,b) < R^+(a-1,b) < R^+(\frac{a+b-1}{2}, \frac{a+b-1}{2})$; while when $a+b$ is an odd number

and $a \leq b-1$, we have $R^+(a,b) < R^+(a,b-1) < R^+(\frac{a+b-1}{2}, \frac{a+b-1}{2})$. Therefore,

$$R^+(a,b) < R^+\left(\lfloor\frac{a+b}{2}\rfloor, \lfloor\frac{a+b}{2}\rfloor\right).$$

According to the second property such that $\lim_{a\to\infty} R^+(a,a) = 0$, we conclude that $\lim_{(a+b)\to\infty} R^+(a,b) = 0$. □

Using Lemma 4.1, we first show that, in any sample path, the optimistic KG will label each instance infinitely many times as T goes to infinity. Let $\eta_i(T)$ be a random variable representing the number of times that the i-th instance has been labeled until the stage T using optimistic KG. Given a sample path ω, let $\mathcal{I}(\omega) = \{i : \lim_{T\to\infty} \eta_i(T)(\omega) < \infty\}$ be the set of instances that has been labeled only finite number of times as T goes to infinity in this sample path. We need to prove that $\mathcal{I}(\omega)$ is an empty set for any ω. We prove it by contradiction. Assuming that $\mathcal{I}(\omega)$ is not empty, then after a certain stage $\widehat{T}$, instances in $\mathcal{I}(\omega)$ will never be labeled. By Lemma 4.1, for any $j \in \mathcal{I}^c$, $\lim_{T\to\infty} R^+(a_j^T(\omega), b_j^T(\omega)) = 0$. Therefore, there will exist $\bar{T} > \widehat{T}$ such that:

$$\begin{aligned}\max_{j\in\mathcal{I}^c} R^+(a_j^{\bar{T}}(\omega), b_j^{\bar{T}}(\omega)) &< \max_{i\in\mathcal{I}} R^+(a_i^{\widehat{T}}(\omega), b_i^{\widehat{T}}(\omega)) \\ &= \max_{i\in\mathcal{I}} R^+(a_i^{\bar{T}}(\omega), b_i^{\bar{T}}(\omega)).\end{aligned}$$

Then according to the optimistic KG policy, the next instance to be labeled must be in $\mathcal{I}(\omega)$, which leads to the contradiction. Therefore, $\mathcal{I}(\omega)$ will be an empty set for any sample path ω.

Let Y_i^s be the random variable which takes the value 1 if the s-th label of the i-th instance is 1 and the value -1 if the s-th label is 0. It is easy to see that $\mathbb{E}(Y_i^s|\theta_i) = \Pr(Y_i^s = 1|\theta_i) = \theta_i$. Hence, Y_i^s, $s = 1, 2, \ldots$ are i.i.d. random variables. By the fact that $\lim_{T\to\infty} \eta_T(i) = \infty$ in all sample paths and using the strong law of large number, we conclude that, conditioning on θ_i, $i = 1, \ldots, K$, the conditional probability of

$$\lim_{T\to\infty} \frac{a_i^T - b_i^T}{\eta_i(T)} = \lim_{T\to\infty} \frac{\sum_{s=1}^{\eta_i(T)} Y_i^s}{\eta_i(T)} = \mathbb{E}(Y_i^s|\theta_i) = 2\theta_i - 1$$

for all $i = 1, \ldots, K$, is one. According to Proposition 2.1., we have $H_T = \{i : a_i^T \geq b_i^T\}$ and $H^* = \{i : \theta_i \geq 0.5\}$. The accuracy is $\mathrm{Acc}(T) = \frac{1}{K}\left(|H_T \cap H^*| + |H_T^c \cap (H^*)^c|\right)$. We have:

$$\begin{aligned}&\Pr(\lim_{T\to\infty} \mathrm{Acc}(T) = 1|\{\theta_i\}_{i=1}^K) \\ =&\Pr\left(\lim_{T\to\infty} (|H_T \cap H^*| + |H_T^c \cap (H^*)^c|) = K|\{\theta_i\}_{i=1}^K\right) \\ \geq&\Pr\left(\lim_{T\to\infty} \frac{a_i^T - b_i^T}{\eta_i(T)} = 2\theta_i - 1, \forall i = 1, \ldots, K|\{\theta_i\}_{i=1}^K\right) \\ =&1,\end{aligned}$$

whenever $\theta_i \neq 0.5$ for all i. The last inequality is due to the fact that, as long as θ_i is not 0.5 in any i, any sample path that gives the event $\lim_{T\to\infty} \frac{a_i^T - b_i^T}{\eta_i(T)} = 2\theta_i - 1, \forall i = 1, \ldots, K$ also gives the event $\lim_{T\to\infty}(a_i^T - b_i^T) = \mathrm{sgn}(2\theta_i - 1)(+\infty)$, which further implies $\lim_{T\to\infty}(|H_T \cap H^*| + |H_T^c \cap (H^*)^c|) = K$. Finally, we have:

$$\begin{aligned}&\Pr\left(\lim_{T\to\infty} \mathrm{Acc}(T) = 1\right) \\ =&\mathbb{E}_{\{\theta_i\}_{i=1}^K}\left[\Pr\left(\lim_{T\to\infty} \mathrm{Acc}(T) = 1|\{\theta_i\}_{i=1}^K\right)\right] \\ =&\mathbb{E}_{\{\theta_i:\theta_i\neq 0.5\}_{i=1}^K}\left[\Pr\left(\lim_{T\to\infty} \mathrm{Acc}(T) = 1|\{\theta_i\}_{i=1}^K\right)\right] \\ =&\mathbb{E}_{\{\theta_i:\theta_i\neq 0.5\}_{i=1}^K}[1] = 1,\end{aligned}$$

where the second equality is because $\{\theta_i : \exists i, \theta_i = 0.5\}$ is a zero measure set.

5. Incorporate Workers' Reliability

We assume there are K instances with the soft-label $\theta_i \sim \mathrm{Beta}(a_i^0, b_i^0)$ and M workers with the reliability $\rho_j \sim \mathrm{Beta}(c_j^0, d_j^0)$. Given the decision on labeling the i-th instance by the j-th worker, we have the probability of the outcome Z_{ij}:

$$\Pr(Z_{ij} = 1|\theta_i, \rho_j) = \theta_i\rho_j + (1-\theta_i)(1-\rho_j) \quad (12)$$

$$\Pr(Z_{ij} = -1|\theta_i, \rho_j) = (1-\theta_i)\rho_j + \theta_i(1-\rho_j) \quad (13)$$

We approximate the posterior so that at any stage for all i, j, θ_i and ρ_j will follow Beta distributions. In particular, assuming at the current state $\theta_i \sim \mathrm{Beta}(a_i, b_i)$ and $\rho_j \sim \mathrm{Beta}(c_j, d_j)$, the posterior distribution conditioned on Z_{ij} takes the following form:

$$p(\theta_i, \rho_j|Z_{ij} = 1) = \frac{\Pr(Z_{ij} = 1|\theta_i, \rho_j)\mathrm{Beta}(a_i, b_i)\mathrm{Beta}(c_j, d_j)}{\Pr(Z_{ij} = 1)}$$

$$p(\theta_i, \rho_j|Z_{ij} = -1) = \frac{\Pr(Z_{ij} = -1|\theta_i, \rho_j)\mathrm{Beta}(a_i, b_i)\mathrm{Beta}(c_j, d_j)}{\Pr(Z_{ij} = -1)}$$

where the likelihood $\Pr(Z_{ij} = z|\theta_i, \rho_j)$ for $z = 1, -1$ is defined in (12) and (13) respectively and

$$\begin{aligned}\Pr(Z_{ij} = 1) &= \mathbb{E}(\Pr(Z_{ij} = 1|\theta_i, \rho_j)) \\ &= \mathbb{E}(\theta_i)\mathbb{E}(\rho_j) + (1 - \mathbb{E}(\theta_i))(1 - \mathbb{E}(\rho_j)) \\ &= \frac{a_i}{a_i + b_i}\frac{c_j}{c_j + d_j} + \frac{b_i}{a_i + b_i}\frac{d_j}{c_j + d_j}.\end{aligned}$$

$$\begin{aligned}\Pr(Z_{ij}=-1) &= \mathbb{E}(\Pr(Z_{ij}=-1|\theta_i,\rho_j))\\ &= (1-\mathbb{E}(\theta_i))\mathbb{E}(\rho_j)+\mathbb{E}(\theta_i)(1-\mathbb{E}(\rho_j))\\ &= \frac{b_i}{a_i+b_i}\frac{c_j}{c_j+d_j}+\frac{a_i}{a_i+b_i}\frac{d_j}{c_j+d_j}.\end{aligned}$$

The posterior distributions $p(\theta_i,p_j|Z_{ij}=z)$ no longer takes the form of the product of Beta distributions on θ_i and p_j. Therefore, we use variational approximation by first assuming the conditional independence of θ_i and ρ_j:

$$p(\theta_i,\rho_j|Z_{ij}=z)\approx p(\theta_i|Z_{ij}=z)p(\rho_j|Z_{ij}=z)$$

In particular, we have the exact form for the marginal distributions:

$$\begin{aligned}p(\theta_i|Z_{ij}=1) &= \frac{\theta_i\mathbb{E}(\rho_j)+(1-\theta_i)(1-\mathbb{E}(\rho_j))}{\Pr(Z_{ij}=1)}\text{Beta}(a_i,b_i)\\ p(\rho_j|Z_{ij}=1) &= \frac{\mathbb{E}(\theta_i)\rho_j+(1-\mathbb{E}(\theta_i))(1-\rho_j)}{\Pr(Z_{ij}=1)}\text{Beta}(c_j,d_j)\\ p(\theta_i|Z_{ij}=-1) &= \frac{(1-\theta_i)\mathbb{E}(\rho_j)+\theta_i(1-\mathbb{E}(\rho_j))}{\Pr(Z_{ij}=-1)}\text{Beta}(a_i,b_i)\\ p(\rho_j|Z_{ij}=-1) &= \frac{(1-\mathbb{E}(\theta_i))\rho_j+\mathbb{E}(\theta_i)(1-\rho_j)}{\Pr(Z_{ij}=-1)}\text{Beta}(c_j,d_j)\end{aligned}$$

To approximate the marginal distribution as Beta distribution, we use the moment matching technique. In particular, we approximate

$$\theta_i|(Z_{ij}=z)\approx\text{Beta}(\tilde{a}_i(z),\tilde{b}_i(z)),$$

such that

$$\widetilde{\mathbb{E}}_z(\theta_i)\doteq\mathbb{E}_{p(\theta_i|Z_{ij}=z)}(\theta_i)=\frac{\tilde{a}_i(z)}{\tilde{a}_i(z)+\tilde{b}_i(z)},\tag{14}$$

$$\widetilde{\mathbb{E}}_z(\theta_i^2)\doteq\mathbb{E}_{p(\theta_i|Z_{ij}=z)}(\theta_i^2)=\frac{\tilde{a}_i(z)(\tilde{a}_i(z)+1)}{(\tilde{a}_i(z)+\tilde{b}_i(z))(\tilde{a}_i(z)+\tilde{b}_i(z)+1)},\tag{15}$$

where $\frac{\tilde{a}_i(z)}{\tilde{a}_i(z)+\tilde{b}_i(z)}$ and $\frac{\tilde{a}_i(z)(\tilde{a}_i(z)+1)}{(\tilde{a}_i(z)+\tilde{b}_i(z))(\tilde{a}_i(z)+\tilde{b}_i(z)+1)}$ are the first and second order moment of $\text{Beta}(\tilde{a}_i(z),\tilde{b}_i(z))$. To make (14) and (15) hold, we have:

$$\tilde{a}_i(z)=\widetilde{\mathbb{E}}_z(\theta_i)\frac{\widetilde{\mathbb{E}}_z(\theta_i)-\widetilde{\mathbb{E}}_z(\theta_i^2)}{\widetilde{\mathbb{E}}_z(\theta_i^2)-\left(\widetilde{\mathbb{E}}_z(\theta_i)\right)^2},\tag{16}$$

$$\tilde{b}_i(z)=(1-\widetilde{\mathbb{E}}_z(\theta_i))\frac{\widetilde{\mathbb{E}}_z(\theta_i)-\widetilde{\mathbb{E}}_z(\theta_i^2)}{\widetilde{\mathbb{E}}_z(\theta_i^2)-\left(\widetilde{\mathbb{E}}_z(\theta_i)\right)^2}.\tag{17}$$

Similarly, we approximate

$$\rho_j|(Z_{ij}=z)\approx\text{Beta}(\tilde{c}_j(z),\tilde{d}_j(z)),$$

such that

$$\widetilde{\mathbb{E}}_z(\rho_j)\doteq\mathbb{E}_{p(\rho_j|Z_{ij}=z)}(\rho_j)=\frac{\tilde{c}_j(z)}{\tilde{c}_j(z)+\tilde{d}_j(z)},\tag{18}$$

$$\widetilde{\mathbb{E}}_z(\rho_j^2)\doteq\mathbb{E}_{p(\rho_j|Z_{ij}=z)}(\rho_j^2)=\frac{\tilde{c}_j(z)(\tilde{c}_j(z)+1)}{(\tilde{c}_j(z)+\tilde{d}_j(z))(\tilde{c}_j(z)+\tilde{d}_j(z)+1)},\tag{19}$$

where $\frac{\tilde{c}_j(z)}{\tilde{c}_j(z)+\tilde{d}_j(z)}$ and $\frac{\tilde{c}_j(z)(\tilde{c}_j(z)+1)}{(\tilde{c}_j(z)+\tilde{d}_j(z))(\tilde{c}_j(z)+\tilde{d}_j(z)+1)}$ are the first and second order moment of $\text{Beta}(\tilde{c}_j(z),\tilde{d}_j(z))$. To make (14) and (15) hold, we have:

$$\tilde{c}_j(z)=\widetilde{\mathbb{E}}_z(\rho_j)\frac{\widetilde{\mathbb{E}}_z(\rho_j)-\widetilde{\mathbb{E}}_z(\rho_j^2)}{\widetilde{\mathbb{E}}_z(\rho_j^2)-\left(\widetilde{\mathbb{E}}_z(\rho_j)\right)^2},\tag{20}$$

$$\tilde{d}_j(z)=(1-\widetilde{\mathbb{E}}_z(\rho_j))\frac{\widetilde{\mathbb{E}}_z(\rho_j)-\widetilde{\mathbb{E}}_z(\rho_j^2)}{\widetilde{\mathbb{E}}_z(\rho_j^2)-\left(\widetilde{\mathbb{E}}_z(\rho_j)\right)^2}.\tag{21}$$

Furthermore, we can compute the exact values for $\widetilde{\mathbb{E}}_z(\theta_i)$, $\widetilde{\mathbb{E}}_z(\theta_i^2)$, $\widetilde{\mathbb{E}}_z(\rho_j)$ and $\widetilde{\mathbb{E}}_z(\rho_j^2)$ as follows.

$$\begin{aligned}\widetilde{\mathbb{E}}_1(\theta_i) &= \frac{\mathbb{E}(\theta_i^2)\mathbb{E}(\rho_j)+(\mathbb{E}(\theta_i)-\mathbb{E}(\theta_i^2))(1-\mathbb{E}(\rho_j))}{p(Z_{ij}=1)}\\ &= \frac{a_i((a_i+1)c_j+b_id_j)}{(a_i+b_i+1)(a_ic_j+b_id_j)}.\end{aligned}$$

$$\begin{aligned}\widetilde{\mathbb{E}}_1(\theta_i^2) &= \frac{\mathbb{E}(\theta_i^3)\mathbb{E}(\rho_j)+(\mathbb{E}(\theta_i^2)-\mathbb{E}(\theta_i^3))(1-\mathbb{E}(\rho_j))}{p(Z_{ij}=1)}\\ &= \frac{a_i(a_i+1)((a_i+2)c_j+b_id_j)}{(a_i+b_i+1)(a_i+b_i+2)(a_ic_j+b_id_j)}.\end{aligned}$$

$$\begin{aligned}\widetilde{\mathbb{E}}_{-1}(\theta_i) &= \frac{(\mathbb{E}(\theta_i)-\mathbb{E}(\theta_i^2))\mathbb{E}(\rho_j)+\mathbb{E}(\theta_i^2)(1-\mathbb{E}(\rho_j))}{p(Z_{ij}=-1)}\\ &= \frac{a_i(b_ic_j+(a_i+1)d_j)}{(a_i+b_i+1)(b_ic_j+a_id_j)}.\end{aligned}$$

$$\begin{aligned}\widetilde{\mathbb{E}}_{-1}(\theta_i^2) &= \frac{(\mathbb{E}(\theta_i^2)-\mathbb{E}(\theta_i^3))\mathbb{E}(\rho_j)+\mathbb{E}(\theta_i^3)(1-\mathbb{E}(\rho_j))}{p(Z_{ij}=-1)}\\ &= \frac{a_i(a_i+1)(b_ic_j+(a_i+2)d_j)}{(a_i+b_i+1)(a_i+b_i+2)(b_ic_j+a_id_j)}.\end{aligned}$$

$$\begin{aligned}\widetilde{\mathbb{E}}_1(\rho_j) &= \frac{\mathbb{E}(\theta_i)\mathbb{E}(\rho_j^2)+(1-\mathbb{E}(\theta_i))(\mathbb{E}(\rho_j)-\mathbb{E}(\rho_j^2))}{p(Z_{ij}=1)}\\ &= \frac{c_j(a_i(c_j+1)+b_id_j)}{(c_j+d_j+1)(a_ic_j+b_id_j)}.\end{aligned}$$

$$\begin{aligned}\widetilde{\mathbb{E}}_1(\rho_j^2) &= \frac{\mathbb{E}(\theta_i)\mathbb{E}(\rho_j^3)+(1-\mathbb{E}(\theta_i))(\mathbb{E}(\rho_j^2)-\mathbb{E}(\rho_j^3))}{p(Z_{ij}=1)}\\ &= \frac{c_j(c_j+1)(a_i(c_j+2)+b_id_j)}{(c_j+d_j+1)(c_j+d_j+2)(a_ic_j+b_id_j)}.\end{aligned}$$

Algorithm 2 Optimistic Knowledge Gradient with Workers' Reliability

Input: Parameters of prior distributions for instances $\{a_i^0, b_i^0\}_{i=1}^K$ and for workers $\{c_j^0, d_j^0\}_{j=1}^M$. The total budget T.

for $t = 0, \ldots, T-1$ **do**

1. Select the next instance i_t to label and the next worker j_t to label i_t according to:

$$(i_t, j_t) = \underset{i\in\{1,\ldots,K\}, j\in\{1,\ldots,M\}}{\arg\max} \left(R^+(a_i^t, b_i^t, c_j^t, d_j^t)\right).$$

Here

$$R^+(a_i^t, b_i^t, c_j^t, d_j^t) = \max(R_1(a_i^t, b_i^t, c_j^t, d_j^t), R_2(a_i^t, b_i^t, c_j^t, d_j^t)).$$

2. Acquire the label $Z_{i_t j_t} \in \{-1, 1\}$ of the i-th instance from the j-th worker.

3. Update the posterior by setting:

$$a_{i_t}^{t+1} = \tilde{a}_{i_t}^t(Z_{i_t j_t}) \qquad b_{i_t}^{t+1} = \tilde{b}_{i_t}^t(Z_{i_t j_t})$$
$$c_{j_t}^{t+1} = \tilde{c}_{j_t}^t(Z_{i_t j_t}) \qquad d_{j_t}^{t+1} = \tilde{d}_{j_t}^t(Z_{i_t j_t}),$$

and all parameters for $i \neq i_t$ and $j \neq j_t$ remain the same.

end for

Output: The positive set $H_T = \{i : a_i^T \geq b_i^T\}$.

$$\widetilde{\mathbb{E}}_{-1}(\rho_j) = \frac{(1-\mathbb{E}(\theta_i))\mathbb{E}(\rho_j^2) + \mathbb{E}(\theta_i)(\mathbb{E}(\rho_j) - \mathbb{E}(\rho_j^2))}{p(Z_{ij} = -1)} = \frac{c_j(b_i(c_j+1) + a_i d_j)}{(c_j + d_j + 1)(b_i c_j + a_i d_j)}.$$

$$\widetilde{\mathbb{E}}_{-1}(\rho_j^2) = \frac{(1-\mathbb{E}(\theta_i))\mathbb{E}(\rho_j^3) + \mathbb{E}(\theta_i)(\mathbb{E}(\rho_j^2) - \mathbb{E}(\rho_j^3))}{p(Z_{ij} = -1)} = \frac{c_j(c_j+1)(b_i(c_j+2) + a_i d_j)}{(c_j + d_j + 1)(c_j + d_j + 2)(b_i c_j + a_i d_j)}.$$

Assuming at a certain stage, θ_i for the i-th instance has the Beta posterior $\text{Beta}(a_i, b_i)$ and ρ_j for the j-th worker has the Beta posterior $\text{Beta}(c_j, d_j)$. The reward of getting label 1 for the i-th instance from the j-th worker and getting label -1 are:

$$R_1(a_i, b_i, c_j, d_j) = h(I(\tilde{a}_i(z=1), \tilde{b}_i(z=1))) - h(I(a_i, b_i)) \quad (22)$$
$$R_2(a_i, b_i, c_j, d_j) = h(I(\tilde{a}_i(z=-1), \tilde{b}_i(z=-1))) - h(I(a_i, b_i)), \quad (23)$$

where $\tilde{a}_i(z = \pm 1)$ and $\tilde{b}_i(z = \pm 1)$ are defined in (20) and (21), which further depend on c_j and d_j. With the reward in place, we present the optimistic knowledge gradient algorithm for budget allocation with the modeling of workers' reliability in Algorithm 2.

6. Extensions

6.1. Incorporating Feature Information

When each instance is associated with a p-dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^p$, we incorporate the feature information in our budget allocation problem by assuming:

$$\theta_i = \sigma(\langle \mathbf{w}, \mathbf{x}_i \rangle) \doteq \frac{1}{1 + \exp\{-\langle \mathbf{w}, \mathbf{x}_i \rangle\}}, \quad (24)$$

where $\sigma(x) = \frac{1}{1+\exp\{-x\}}$ is the sigmoid function and $\mathbf{w}$ is assumed to be drawn from a Gaussian prior $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. At the t-th stage with the state $S^t = (\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ and $\mathbf{w} \sim (\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$, one decides to label the i_t-th instance according to a certain policy (e.g., KG) and observes the label $y_{i_t} \in \{-1, 1\}$. The posterior distribution $p(\mathbf{w}|y_{i_t}, S^t) \propto p(y_{i_t}|\mathbf{w})p(\mathbf{w}|S^t)$ has the following log-likelihood:

$$\begin{aligned}
&\ln p(\mathbf{w}|y_{i_t}, S^t) \\
=&\ln p(y_{i_t}|\mathbf{w}) + \ln p(\mathbf{w}|S^t) + \text{const} \\
=&\mathbf{1}(y_{i_t} = 1) \ln \sigma(\langle \mathbf{w}, \mathbf{x}_{i_t} \rangle) + \mathbf{1}(y_{i_t} = -1) \ln (1 - \sigma(\langle \mathbf{w}, \mathbf{x}_{i_t} \rangle)) \\
&- \frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_t)' \boldsymbol{\Omega}_t (\mathbf{w} - \boldsymbol{\mu}_t) + \text{const},
\end{aligned}$$

where $\boldsymbol{\Omega}_t = (\boldsymbol{\Sigma}_t)^{-1}$ is the precision matrix. To approximate $p(\mathbf{w}|y_{i_t}, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)$ by a Gaussian distribution $N(\boldsymbol{\mu}_{t+1}, \boldsymbol{\Sigma}_{t+1})$, we use the Laplace method (see Chapter 4.4 in (Bishop, 2007)). In particular, we define the mean of the posterior Gaussian using the MAP (maximum a posteriori) estimator of $\mathbf{w}$:

$$\mu_{t+1} = \arg\max_{\mathbf{w}} \ln p(\mathbf{w}|y_{i_t}, S^t). \quad (25)$$

And μ_{t+1} can be solved by Newton's method. and the precision matrix:

$$\begin{aligned}
\boldsymbol{\Omega}_{t+1} &= -\nabla^2 \ln p(\mathbf{w}|y_{i_t}, S^t)\big|_{\mathbf{w}=\boldsymbol{\mu}_{t+1}} \\
&= \boldsymbol{\Omega}_t + \sigma(\boldsymbol{\mu}'_{t+1}\mathbf{x}_{i_{t+1}})(1 - \sigma(\boldsymbol{\mu}'_{t+1}\mathbf{x}_{i_{t+1}}))\mathbf{x}_{i_{t+1}}\mathbf{x}'_{i_{t+1}}.
\end{aligned}$$

By Sherman-Morrison formula, the covariance matrix

$$\begin{aligned}
&\boldsymbol{\Sigma}_{t+1} = (\boldsymbol{\Omega}_{t+1})^{-1} \\
=&\boldsymbol{\Sigma}_t - \frac{\sigma(\boldsymbol{\mu}'_{t+1}\mathbf{x}_{i_t})(1 - \sigma(\boldsymbol{\mu}_{t+1}\mathbf{x}_{i_t}))}{1 + \sigma(\boldsymbol{\mu}'_{t+1}\mathbf{x}_{i_t})(1 - \sigma(\boldsymbol{\mu}'_{t+1}\mathbf{x}_{i_t}))\mathbf{x}'_{i_t}\boldsymbol{\Sigma}_t\mathbf{x}_{i_t}} \boldsymbol{\Sigma}_t \mathbf{x}_{i_{t+1}} \mathbf{x}'_{i_t} \boldsymbol{\Sigma}_t.
\end{aligned}$$

We also calculate the transition probability of $y_{i_t} = 1$ and $y_{i_t} = -1$ as follows:

$$\begin{aligned}
\Pr(y_{i_t} = 1|S^t, i_t) &= \int p(y_{i_t} = 1|\mathbf{w}) p(\mathbf{w}|S^t) \mathrm{d}\mathbf{w} \\
&= \int \sigma(\mathbf{w}'\mathbf{x}_i) p(\mathbf{w}|S^t) \mathrm{d}\mathbf{w} \\
&\approx \sigma(\mu_i \kappa(s_i^2)),
\end{aligned}$$

where $\kappa(s_i^2) = (1 + \pi s_i^2/8)^{-1/2}$ and $\mu_i = \langle \boldsymbol{\mu}_t, \mathbf{x}_i \rangle$ and $s_i^2 = \mathbf{x}_i' \boldsymbol{\Sigma}_t \mathbf{x}_i$.

To calculate the reward function, in addition to the transition probability, we also need to compute:

$$\begin{aligned} P_i^t &= \Pr(\theta_i \geq 0.5 | \mathcal{F}_t) \\ &= \Pr\left(\frac{1}{1+\exp\{-\mathbf{w}_t' \mathbf{x}_i\}} \geq 0.5 \Big| \mathbf{w}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \right) \\ &= \Pr(\mathbf{w}_t' \mathbf{x}_i \geq 0 | \mathbf{w}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)) \\ &= \int_0^\infty \left(\int_{\mathbf{w}} \delta(c - \langle \mathbf{w}, \mathbf{x}_i \rangle) N(\mathbf{w} | \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \mathrm{d}\mathbf{w} \right) \mathrm{d}c, \end{aligned}$$

where $\delta(\cdot)$ is the Dirac delta function.Let

$$p(c) = \int_{\mathbf{w}} \delta(c - \langle \mathbf{w}, \mathbf{x}_i \rangle) N(\mathbf{w} | \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \mathrm{d}\mathbf{w}.$$

Since the marginal of a Gaussian distribution is still a Gaussian, $p(c)$ is a univariate-Gaussian distribution with the mean and variance:

$$\begin{aligned} \mu_i &= \mathbb{E}(c) = \langle \mathbb{E}(\mathbf{w}), \mathbf{x}_i \rangle = \langle \boldsymbol{\mu}_t, \mathbf{x}_i \rangle \\ s_i^2 &= \mathrm{Var}(c) = (\mathbf{x}_i)' \mathrm{Cov}(\mathbf{w}, \mathbf{w}) \mathbf{x}_i = (\mathbf{x}_i)' \boldsymbol{\Sigma}_t \mathbf{x}_i. \end{aligned}$$

Therefore, we have:

$$P_i^t = \int_0^\infty p(c) \mathrm{d}c = 1 - \Phi\left(-\frac{\mu_i}{s_i} \right), \tag{26}$$

where $\Phi(\cdot)$ is the Gaussian CDF.

With P_i^t and transition probability in place, the expected reward in value function takes the following form :

$$R(S^t, i_t) = \mathbb{E}\left(\sum_{i=1}^K h(P_i^{t+1}) - \sum_{i=1}^K h(P_i^t) \Big| S^t, i_t \right). \tag{27}$$

We note that since $\mathbf{w}$ will affect all P_i^t, the summation from 1 to K in (27) can not be omitted and hence (27) cannot be written as $\mathbb{E}\left(h(P_{i_t}^{t+1}) - h(P_{i_t}^t) | S^t, i_t\right)$ in (3). In this problem, the myopic KG or optimistic KG need to solve $O(2TK)$ optimization problems to compute the mean of the posterior as in (25), which could be computationally quite expensive. One possibility to address this problem is to use the variational Bayesian logistic regression (Jaakkola & Jordan, 2000), which could lead to a faster optimization procedure.

6.2. Multi-Class Setting

In multi-class setting with C different classes, we assume that the i-th instance is associated with a probability vector $\boldsymbol{\theta}_i = (\theta_{i1}, \dots \theta_{iC})$, where θ_{ic} is the probability that the i-th instance belongs to the class c and $\sum_{i=1}^C \theta_{ic} = 1$. We assume that $\boldsymbol{\theta}_i$ has a Dirichlet prior $\boldsymbol{\theta}_i \sim \mathrm{Dir}(\boldsymbol{\alpha}_i^0)$ and our initial state S^0 is a $K \times C$ matrix with $\boldsymbol{\alpha}_i^0$ as its i-th row. At each stage t with the current state S^t, we determine an instance i_t to label and collect its label $y_{i_t} \in \{1, \dots, C\}$, which follows the categorical distribution: $p(y_{i_t}) = \prod_{c=1}^C \theta_{i_t c}^{I(y_{i_t}=c)}$. Since the Dirichlet is the conjugate prior of the categorical distribution, the next state induced by the posterior distribution is: $S_{i_t}^{t+1} = S_{i_t}^t + \boldsymbol{\delta}_{y_{i_t}}$ and $S_i^{t+1} = S_i^t$ for all $i \neq i_t$. Here $\boldsymbol{\delta}_c$ is a row vector with one at the c-th entry and zeros at all other entries. The transition probability:

$$\Pr(y_{i_t} = c | S^t, i_t) = \mathbb{E}(\theta_{i_t c} | S^t) = \frac{\alpha_{i_t c}^t}{\sum_{c=1}^C \alpha_{i_t c}^t}.$$

In multi-class problem, at the final stage T when all budget is used up, we construct the set H_c^T for each class c to maximize the conditional expected classification accuracy:

$$\begin{aligned} \{H_c^T\}_{c=1}^C &= \underset{H_c \subseteq \{1,\dots,C\}, H_c \cap H_{\tilde{c}} = \emptyset}{\arg\max} \mathbb{E}\left(\sum_{i=1}^K \sum_{c=1}^C I(i \in H_c) I(i \in H_c^*) \Big| \mathcal{F}_T \right) \\ &= \underset{H_c \subseteq \{1,\dots,C\}, H_c \cap H_{\tilde{c}} = \emptyset}{\arg\max} \sum_{i=1}^K \sum_{c=1}^C I(i \in H_c) \Pr\left(i \in H_c^* | \mathcal{F}_T\right). \end{aligned} \tag{28}$$

Here, $H_c^* = \{i : \theta_{ic} \geq \theta_{ic'}, \forall c' \neq c\}$ is the true set of instances in the class c. The set H_c^T consists of instances that belong to class c. Therefore, $\{H_c^T\}_{c=1}^C$ should form a partition of all instances $\{1, \dots, K\}$. Let

$$P_{ic}^T = \Pr(i \in H_c^* | \mathcal{F}_T) = \Pr(\theta_{ic} \geq \theta_{i\tilde{c}}, \ \forall \ \tilde{c} \neq c | \mathcal{F}_T). \tag{29}$$

To maximize RHS of (28), we have

$$H_c^T = \{i : P_{ic}^T \geq P_{i\tilde{c}}^T, \forall \tilde{c} \neq c\}. \tag{30}$$

If there is i belongs to more than one H_c^T, we only assign it to the one with the smallest index c. The maximum conditional expected accuracy takes the form:

$$\sum_{i=1}^K \left(\max_{c \in \{1\dots,C\}} P_{ic}^T \right). \tag{31}$$

Then the value function can be defined as:

$$\begin{aligned} V(S^0) &\doteq \sup_\pi \mathbb{E}^\pi \left(\sum_{i=1}^K \sum_{c=1}^C I(i \in H_c^T) I(i \in H_c^*) \Big| S^0 \right) \\ &= \sup_\pi \mathbb{E}^\pi \left(\mathbb{E}^\pi \left(\sum_{i=1}^K \sum_{c=1}^C I(i \in H_c^T) I(i \in H_c^*) \Big| \mathcal{F}_T \right) \Big| S^0 \right) \\ &= \sup_\pi \mathbb{E}^\pi \left(\sum_{i=1}^K h(\mathbf{P}_i^T) \Big| S^0 \right), \end{aligned} \tag{32}$$

where $\mathbf{P}_i^T = (P_{i1}^T, \dots, P_{iC}^T)$ and

$$h(\mathbf{P}_i^T) \doteq \max_{c \in \{1\dots,C\}} P_{ic}^T.$$

Following Proposition 2.2, let $P_{ic}^t = \Pr(i \in H_c^* | \mathcal{F}_t)$ and $\mathbf{P}_i^t = (P_{i1}^t, \ldots, P_{iC}^t)$. By defining intermediate reward function at each stage:

$$R(S^t, i_t) = \mathbb{E}\left(h(\mathbf{P}_{i_t}^{t+1}) - h(\mathbf{P}_{i_t}^t) | S^t, i_t\right).$$

The value function can be re-written as:

$$V(S^0) = G_0(S^0) + \sup_{\pi} \mathbb{E}^{\pi}\left(\sum_{t=0}^{T-1} R(S^t, i_t) \Big| S^0\right),$$

where $G_0(S^0) = \sum_{i=1}^K h(\mathbf{P}_i^0)$. Since the reward function only depends on $S_{i_t}^t = \boldsymbol{\alpha}_{i_t}^t \in \mathbb{R}_+^C$, we can define the reward function in a more explicit way by defining:

$$R(\boldsymbol{\alpha}) = \sum_{c=1}^{C} \frac{\alpha_c}{\sum_{\tilde{c}=1}^C \alpha_{\tilde{c}}} h(I(\boldsymbol{\alpha} + \boldsymbol{\delta}_c)) - h(I(\boldsymbol{\alpha})).$$

Here $\boldsymbol{\delta}_c$ be a row vector of length C with one at the c-th entry and zeros at all other entries; and $I(\boldsymbol{\alpha}) = (I_1(\boldsymbol{\alpha}), \ldots, I_C(\boldsymbol{\alpha}))$ where

$$I_c(\boldsymbol{\alpha}) = \Pr(\theta_c \geq \theta_{\tilde{c}}, \forall \tilde{c} \neq c | \theta \sim \mathrm{Dir}(\boldsymbol{\alpha})). \tag{33}$$

Therefore, we have $R(S^t, i_t) = R(\boldsymbol{\alpha}_{i_t}^t)$.

To evaluate the reward $R(\boldsymbol{\alpha})$, the major bottleneck is how to compute $I_c(\boldsymbol{\alpha})$ efficiently. Directly taking the C-dimensional integration on the region $\{\theta_c \geq \theta_{\tilde{c}}, \forall \tilde{c} \neq c\} \cap \Delta_C$ will be computationally very expensive, where Δ_C denotes the C-dimensional simplex. Therefore, we propose a method to convert the computation of $I_c(\boldsymbol{\alpha})$ into a one-dimensional integration. It is known that to generate $\theta \sim \mathrm{Dir}(\boldsymbol{\alpha})$, it is equivalent to generate $\{X_c\}_{c=1}^C$ with $X_c \sim \mathrm{Gamma}(\alpha_c, 1)$ and let $\theta_c \equiv \frac{X_c}{\sum_{c=1}^C X_c}$. Then $\theta = (\theta_1, \ldots, \theta_C)$ will follow $\mathrm{Dir}(\boldsymbol{\alpha})$. Therefore, we have:

$$I_c(\boldsymbol{\alpha}) = \Pr(X_c \geq X_{\tilde{c}}, \forall \tilde{c} \neq c | X_c \sim \mathrm{Gamma}(\alpha_c, 1)). \tag{34}$$

It is easy to see that

$$\begin{aligned} & I_c(\boldsymbol{\alpha}) \qquad\qquad (35) \\ &= \int_{0 \leq x_1 \leq x_c} \cdots \int_{x_c \geq 0} \cdots \int_{0 \leq x_C \leq x_c} \prod_{c=1}^{C} f_{\mathrm{Gamma}}(x_c; \alpha_c, 1) \mathrm{d}x_1 \ldots \mathrm{d}x_C \\ &= \int_{x_c \geq 0} f_{\mathrm{Gamma}}(x_c; \alpha_c, 1) \prod_{\tilde{c} \neq c} F_{\mathrm{Gamma}}(x_c; \alpha_{\tilde{c}}, 1) \mathrm{d}x_c, \end{aligned}$$

where $f_{\mathrm{Gamma}}(x; \alpha_c, 1)$ is the density function of Gamma distribution with the parameter $(\alpha_c, 1)$ and $F_{\mathrm{Gamma}}(x_c; \alpha_{\tilde{c}}, 1)$ is the CDF of Gamma distribution at x_c with the parameter $(\alpha_{\tilde{c}}, 1)$. In many softwares, $F_{\mathrm{Gamma}}(x_c; \alpha_{\tilde{c}}, 1)$ can be calculated very efficiently without an explicit integration. Therefore, we can evaluate $I_c(\boldsymbol{\alpha})$ by performing only a one-dimensional numerical integration as in (35). We could also use Monte-Carlo approximation to further accelerate the computation in (35).

References

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2007.

Jaakkola, T. and Jordan, M. I. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

Xie, J. and Frazier, P. I. Sequential bayes-optimal policies for multiple comparions with a control. Technical report, Cornell University, 2012.