A. Discussion of Assumptions

In this section, we prove that the combinatorial semi-bandit and the cascading bandit satisfy Assumptions 1 and 2 proposed in Section 5.

A.1. Combinatorial Semi-Bands

Notice that in a combinatorial semi-bandit, the action $a = (a_1, \ldots, a_K)$, and

$$r(a, \theta) = \sum_{k=1}^{K} \theta^{(a_k)} = \sum_{l=1}^{L} \theta^{(l)}1_{l \in a}.$$ 

Thus, for any $l$, $r(a, \theta)$ is weakly increasing in $\theta^{(l)}$. Hence Assumption 1 is satisfied. On the other hand, we have

$$|r(a, \theta_1) - r(a, \theta_2)| = \left| \sum_{l=1}^{L} \left( \theta^{(l)}_1 - \theta^{(l)}_2 \right) 1_{l \in a} \right|$$

$$\leq \sum_{l=1}^{L} \left| \theta^{(l)}_1 - \theta^{(l)}_2 \right| 1_{l \in a} = \sum_{l=1}^{L} P\left( E^{(l)} \mid \theta_2, a \right) \left| \theta^{(l)}_1 - \theta^{(l)}_2 \right|,$$

where the last quality follows from the fact that all nodes in a combinatorial semi-bandit is observed, and hence $P\left( E^{(l)} \mid \theta, a \right) = 1_{l \in a}$ for all $\theta$. Thus, Assumption 2 is satisfied with $C = 1$.

A.2. Cascading Bandits

For a cascading bandit, the action is $a = (a_1, \ldots, a_K)$, and

$$r(a, \theta) = 1 - \prod_{k=1}^{K} (1 - \theta^{(a_k)}) = 1 - \prod_{l \in a} (1 - \theta^{(l)}).$$

Thus, for any $l$, $r(a, \theta)$ is weakly increasing in $\theta^{(l)}$. Hence Assumption 1 is satisfied. On the other hand, from Kveton et al. (2015a), we have

$$r(a_1, \theta) - r(a_2, \theta) = \sum_{k=1}^{K} \prod_{k=1}^{K} (1 - \theta^{(a_k)}) \left( \theta^{(a_k)}_1 - \theta^{(a_k)}_2 \right) \prod_{k_2=k+1}^{K} (1 - \theta^{(a_{k_2})}_1)$$

$$= \sum_{k=1}^{K} P\left( E^{(a_k)} \mid \theta_2, a \right) \left( \theta^{(a_k)}_1 - \theta^{(a_k)}_2 \right) \prod_{k_2=k+1}^{K} (1 - \theta^{(a_{k_2})}_1),$$

where the second equality follows from $P\left( E^{(a_k)} \mid \theta_2, a \right) = \prod_{k_1=1}^{K} (1 - \theta^{(a_{k_1})}_2)$. Thus, we have

$$|r(a_1, \theta) - r(a_2, \theta)| = \left| \sum_{k=1}^{K} P\left( E^{(a_k)} \mid \theta_2, a \right) \left( \theta^{(a_k)}_1 - \theta^{(a_k)}_2 \right) \prod_{k_2=k+1}^{K} (1 - \theta^{(a_{k_2})}_1) \right|$$

$$\leq \sum_{k=1}^{K} P\left( E^{(a_k)} \mid \theta_2, a \right) \left| \theta^{(a_k)}_1 - \theta^{(a_k)}_2 \right| \prod_{k_2=k+1}^{K} (1 - \theta^{(a_{k_2})}_1)$$

$$\leq \sum_{k=1}^{K} P\left( E^{(a_k)} \mid \theta_2, a \right) \left| \theta^{(a_k)}_1 - \theta^{(a_k)}_2 \right| \prod_{k_2=k+1}^{K} (1 - \theta^{(a_{k_2})}_1),$$

where the last inequality follows from $\prod_{k_2=k+1}^{K} (1 - \theta^{(a_{k_2})}_1) \in [0, 1]$. Thus, Assumption 2 is satisfied with $C = 1$. 

Influence Diagram Bandits: Variational Thompson Sampling for Structured Bandit Problems
B. Proof for Theorem 1

Proof:
Recall that the stochastic instantaneous reward is \( r(x, z) \). Note that \( r(x, z) \) is bounded since its domain is finite. Without loss of generality, we assume that \( r(x, z) \in [0, B] \). Thus, for any action \( a \) and probability measure \( \theta \in [0, 1]^{d+L} \), we have \( r(a, \theta) \in [0, B] \).

Define \( R_t = r(a^*, \theta_a) - r(a_t, \theta_a) \), then by definition, we have

\[
R_B(n) = \sum_{t=1}^{n} \mathbb{E}[R_t] = \sum_{t=1}^{n} \mathbb{E}[E[R_t|\mathcal{H}_{t-1}]],
\]

where \( \mathcal{H}_{t-1} \) is the “history” by the end of time \( t - 1 \), which includes all the actions and observations by that time\(^5\). For any parameter index \( i = 1, \ldots, d + L \) and any time \( t \), we define \( N_t^{(i)} = \sum_{r=1}^{t} 1 \{ E_r^{(i)} \} \) as the number of times that the samples corresponding to parameter \( \theta_r^{(i)} \) have been observed by the end of time \( t \), and \( \hat{\theta}_r^{(i)} \) as the empirical mean for \( \theta_r^{(i)} \) based on these \( N_t^{(i)} \) observations. Then we define the upper confidence bound (UCB) \( U_t^{(i)} \) and the lower confidence bound (LCB) \( L_t^{(i)} \) as

\[
U_t^{(i)} = \begin{cases} 
\min \left\{ \hat{\theta}_t^{(i)} + c(t, N_t^{(i)}), 1 \right\} & \text{if } N_t^{(i)} > 0 \\
1 & \text{otherwise}
\end{cases}, \quad L_t^{(i)} = \begin{cases} 
\max \left\{ \hat{\theta}_t^{(i)} - c(t, N_t^{(i)}), 0 \right\} & \text{if } N_t^{(i)} > 0 \\
0 & \text{otherwise}
\end{cases}
\]

where \( c(t, N) = \sqrt{\frac{\text{log}(t)}{N}} \) for any positive integer \( t \) and \( N \). Moreover, we define a probability measure \( \hat{\theta}_t \in [0, 1]^{d+L} \) as

\[
\hat{\theta}_t^{(i)} = \begin{cases} 
U_t^{(i)} & \text{if } i \in \mathcal{I}^+ \\
L_t^{(i)} & \text{if } i \in \mathcal{I}^-
\end{cases}
\]

Since both \( N_{t-1}^{(i)} \) and \( \hat{\theta}_t^{(i)} \) are conditionally deterministic given \( \mathcal{H}_{t-1} \), and \( \mathcal{I}^+ \) and \( \mathcal{I}^- \) are deterministic, by the definitions above, \( U_{t-1} \), \( L_{t-1} \) and \( \theta_{t-1} \) are also conditionally deterministic given \( \mathcal{H}_{t-1} \). Moreover, as is discussed in Russo & Van Roy (2014), since we apply exact Thompson sampling i.d.TS, \( \theta_t \) and \( \hat{\theta}_t \) are conditionally i.i.d. given \( \mathcal{H}_{t-1} \), and \( a^* = \arg \max_a r(a, \theta_a) \) and \( a_t = \arg \max_a r(a, \theta_t) \). Thus, conditioning on \( \mathcal{H}_{t-1} \), \( r(a^*, \theta_t) \) and \( r(a_t, \theta_t) \) are i.i.d., consequently, we have

\[
\mathbb{E}[R_t|\mathcal{H}_{t-1}] = \mathbb{E}[r(a^*, \theta_a) - r(a_t, \theta_a)|\mathcal{H}_{t-1}] \\
= \mathbb{E}[r(a^*, \theta_a) - r(a^*, \theta_{t-1})|\mathcal{H}_{t-1}] + \mathbb{E}[r(a_t, \theta_{t-1}) - r(a_t, \theta_a)|\mathcal{H}_{t-1}].
\]

To simplify the exposition, for any time \( t \) and \( i = 1, \ldots, d + L \), we define

\[
G_t^{(i)} = \left\{ \left| \theta_r^{(i)} - \hat{\theta}_t^{(i)} \right| > c(t, N_t^{(i)}), N_t^{(i)} > 0 \right\} = \left\{ \theta_r^{(i)} > U_t^{(i)} \text{ or } \theta_r^{(i)} < L_t^{(i)} \right\}.
\]

Notice that \( \bigcup_{i=1}^{d+L} G_t^{(i)} = \bigcup_{i=1}^{d+L} G_t^{(i)} \{ L_t \leq \theta_s \leq U_t \} \). Moreover, from Assumption 1, if \( L_t \leq \theta_s \leq U_t \), based on the definition of \( \theta_t \), we have \( r(a, \theta_s) \leq r(a, \theta_t) \) for all action \( a \). Thus, we have

\[
r(a^*, \theta_s) - r(a^*, \theta_{t-1}) \overset{(a)}{=} [r(a^*, \theta_s) - r(a^*, \theta_{t-1})] 1 \{ L_{t-1} \leq \theta_s \leq U_{t-1} \} \\
\overset{(b)}{\leq} [r(a^*, \theta_s) - r(a^*, \theta_{t-1})] \left( \bigcup_{i=1}^{d+L} G_t^{(i)} \right) \\
\overset{(c)}{\leq} B1 \left( \bigcup_{i=1}^{d+L} G_t^{(i)} \right) \overset{(d)}{\leq} B \sum_{i=1}^{d+L} 1 \{ G_t^{(i)} \}.
\]

\(^5\)Rigorously speaking, \( \{ \mathcal{H}_i \}_{i=0}^{\infty} \) is a filtration and \( \mathcal{H}_{t-1} \) is a \( \sigma \)-algebra.
where equality (a) is simply a decomposition based on indicators, inequality (b) follows from the fact that \( r(a, \theta_s) \leq r(a, \theta_{t-1}) \) if \( L_{t-1} \leq \theta_s \leq U_{t-1} \), inequality (c) follows from the fact that \( r(X, Z) \in [0, B] \) for all \((X, Z)\) and hence \( r(a, \theta) \in [0, B] \) for all \(a\) and \(\theta\), and inequality (d) trivially follows from the union bound of the indicators.

On the other hand, we have
\[
\begin{align*}
\left( a \right) & \Rightarrow r(a_t, \theta_{t-1}) - r(a_t, \theta_s) = [r(a_t, \theta_{t-1}) - r(a_t, \theta_s)] \mathbf{1} (L_{t-1} \leq \theta_s \leq U_{t-1}) \\
& \quad + [r(a_t, \theta_{t-1}) - r(a_t, \theta_s)] \mathbf{1} \left( \bigcup_{i=1}^{d+L} G_{t-1}^{(i)} \right).
\end{align*}
\]

Similarly as the above analysis, we have
\[
[r(a_t, \theta_{t-1}) - r(a_t, \theta_s)] \mathbf{1} \left( \bigcup_{i=1}^{d+L} G_{t-1}^{(i)} \right) \leq B \sum_{i=1}^{d+L} \mathbf{1} \left( G_{t-1}^{(i)} \right).
\]

On the other hand, we have
\[
[r(a_t, \theta_{t-1}) - r(a_t, \theta_s)] \mathbf{1} (L_{t-1} \leq \theta_s \leq U_{t-1}) \leq C \sum_{i=1}^{d+L} \mathbf{E} \left[ P \left( G_{t-1}^{(i)} \right) \left| H_{t-1} \right. \right] + 2B \sum_{i=1}^{d+L} \mathbf{E} \left[ \mathbf{1} \left( G_{t-1}^{(i)} \right) \right] + 2B \sum_{i=1}^{d+L} \mathbf{E} \left[ \mathbf{1} \left( G_{t-1}^{(i)} \right) \right].
\]

where inequality (a) follows from Assumption 2, inequality (b) follows trivially from \( L_{t-1} \leq \theta_s \leq U_{t-1} \) and the definition of \( \theta_{t-1} \), and inequality (c) follows from the fact that \( U_{t-1} > L_{t-1} \) always holds, no matter what \(\theta_s\) is. Combining the above results, we have
\[
\mathbf{E}[R_t \mid H_{t-1}] \leq C \sum_{i=1}^{d+L} \mathbf{E} \left[ P \left( G_{t-1}^{(i)} \right) \left| H_{t-1} \right. \right] + 2B \sum_{i=1}^{d+L} \mathbf{E} \left[ \mathbf{1} \left( G_{t-1}^{(i)} \right) \right] + 2B \sum_{i=1}^{d+L} \mathbf{E} \left[ \mathbf{1} \left( G_{t-1}^{(i)} \right) \right].
\]

where (a) follows from the fact that \( U_{t-1} \) and \( L_{t-1} \) are deterministic conditioning on \( H_{t-1} \), (b) follows from the definition of \( P \left( G_{t-1}^{(i)} \right) \), (c) follows from that fact that conditioning on \( \theta_s \) and \( a_t, E_t^{(i)} \) is independent of \( H_{t-1} \), and (d) follows from the tower property. Thus we have
\[
R_B(n) \leq C \sum_{i=1}^{d+L} \sum_{t=1}^{n} \mathbf{E} \left[ P \left( G_{t-1}^{(i)} \right) \left| H_{t-1} \right. \right] + 2B \sum_{i=1}^{d+L} \sum_{t=1}^{n} P \left( G_{t-1}^{(i)} \right).
\]

We first bound the second term. Notice that we have \( P \left( G_{t-1}^{(i)} \right) = \mathbf{E} \left[ P \left( G_{t-1}^{(i)} \right) \left| \theta_s \right. \right] \). For any \( \theta_s \), we have
\[
P \left( G_{t-1}^{(i)} \left| \theta_s \right. \right) = P \left( \theta_s - \theta_s^{(i)}_{t-1} \right) > c (t, \theta_s^{(i)}_{t-1}, \theta_s^{(i)}_{t-1} > 0 \left| \theta_s \right.).
\]
where we use subscript $N_{t-1}^{(i)}$ for $\hat{\theta}$ to emphasize it is an empirical mean over $N_{t-1}^{(i)}$ samples. Following the union bound developed in Auer et al. (2002), we have

$$P\left(G_{t-1}^{(i)}|\theta_*\right) = P\left(\left|\theta_{t-1}^{(i)} - \hat{\theta}_{N_{t-1}^{(i)}}^{(i)}\right| > c\left(t, N_{t-1}^{(i)}\right), N_{t-1}^{(i)} > 0 | \theta_*\right)$$

$$\leq \sum_{n=1}^{d+L} P\left(\left|\theta_{t-1}^{(i)} - \hat{\theta}_{N_{t-1}^{(i)}}^{(i)}\right| > c\left(t, N\right) | \theta_*\right) \leq \sum_{t=1}^{N-1} \frac{2}{t^2} < \frac{2}{L^2},$$

where inequality (a) follows from the union bound over the realization of $N_{t-1}^{(i)}$, and inequality (b) follows from the Hoeffding’s inequality. Since the above inequality holds for any $\theta_*$, we have $P\left(G_{t-1}^{(i)}\right) < \frac{2}{L^2}$. Thus,

$$\sum_{i=1}^{d+L} \sum_{t=1}^{n} P \left( G_{t-1}^{(i)} \right) < \sum_{i=1}^{d+L} \sum_{t=1}^{n} \frac{2}{t^2} < \left( d + L \right) \frac{2}{L^2} = \frac{(d + L)^2}{3}.$$  

We now try to bound the first term of equation 12. Notice that trivially, we have

$$U_{t-1}^{(i)} - L_{t-1}^{(i)} \leq 2c\left(t, N_{t-1}^{(i)}\right) 1\left(N_{t-1}^{(i)} > 0\right) + 1\left(N_{t-1}^{(i)} = 0\right)$$

$$= 2 \sqrt{1.5 \log(t) \frac{n}{N_{t-1}^{(i)}}} 1\left(N_{t-1}^{(i)} > 0\right) + 1\left(N_{t-1}^{(i)} = 0\right)$$

$$\leq \sqrt{6 \log(n)} \frac{1}{\sqrt{N_{t-1}^{(i)}}} 1\left(N_{t-1}^{(i)} > 0\right) + 1\left(N_{t-1}^{(i)} = 0\right).$$

Thus, we have

$$\sum_{i=1}^{d+L} \sum_{t=1}^{n} \frac{1}{\sqrt{N_{t-1}^{(i)}}} 1\left(E_{t}^{(i)}, N_{t-1}^{(i)} > 0\right) \leq \sqrt{6 \log(n)} \sum_{i=1}^{d+L} \sum_{t=1}^{n} \frac{1}{\sqrt{N_{t-1}^{(i)}}} 1\left(E_{t}^{(i)}, N_{t-1}^{(i)} > 0\right) + (d + L).$$

Notice that from the Cauchy–Schwarz inequality, we have

$$\sum_{i=1}^{d+L} \sum_{t=1}^{n} \frac{1}{\sqrt{N_{t-1}^{(i)}}} 1\left(E_{t}^{(i)}, N_{t-1}^{(i)} > 0\right) \leq \sqrt{\sum_{i=1}^{n} \sum_{t=1}^{d+L} 1\left(E_{t}^{(i)}\right)} \sum_{i=1}^{d+L} \sum_{t=1}^{n} \frac{1}{\sqrt{N_{t-1}^{(i)}}} 1\left(N_{t-1}^{(i)} > 0\right).$$

Moreover, we have

$$\sum_{i=1}^{d+L} \sum_{t=1}^{n} \frac{1}{\sqrt{N_{t-1}^{(i)}}} 1\left(N_{t-1}^{(i)} > 0\right) < (d + L) \sum_{N=1}^{n} \frac{1}{N} < (d + L) \left(1 + \int_{z=1}^{n} \frac{1}{z} dz\right) = (d + L)(1 + \log(n)).$$

Consequently, we have

$$\mathbb{E}\left[ \sum_{i=1}^{d+L} \sum_{t=1}^{n} 1\left(E_{t}^{(i)}\right)\left[U_{t-1}^{(i)} - L_{t-1}^{(i)}\right]\right] \leq \sqrt{6(d + L) \log(n) \left(1 + \log(n)\right) \mathbb{E}\left[ \sum_{t=1}^{n} \sum_{i=1}^{d+L} 1\left(E_{t}^{(i)}\right)\right] + (d + L).$$

Moreover, we have

$$\mathbb{E}\left[ \sqrt{\sum_{t=1}^{n} \sum_{i=1}^{d+L} 1\left(E_{t}^{(i)}\right)}\right] \leq \sqrt{\sum_{t=1}^{n} \mathbb{E}\left[ \sum_{i=1}^{d+L} 1\left(E_{t}^{(i)}\right)\right]} \leq \sqrt{n} \mathbb{E}\left[ \max_{a} \left[ \sum_{i=1}^{d+L} 1\left(E_{t}^{(i)}\right)\right] a\right]$$

$$\leq \sqrt{n} \mathbb{E}\left[ \max_{a} \left[ \sum_{t=1}^{d+L} 1\left(E_{t}^{(i)}\right)\right] a\right] \leq \sqrt{n} \mathbb{E}[O_{\text{max}}] = \sqrt{nO_{\text{max}}},$$
where equality (a) follows from the tower property, and equality (b) follows from the definition of $O_{\text{max}}$. Thus, we have

$$\sum_{i=1}^{d+L} \sum_{t=1}^{n} 1 \left( E_{t}^{(i)} - U_{t-1}^{(i)} \right) \leq \sqrt{6(d+L)} O_{\text{max}} n \log(n) (1 + \log(n)) + (d + L)$$

Putting everything together, we have

$$R_B(n) \leq C \sqrt{6(d+L)} O_{\text{max}} n \log(n) (1 + \log(n)) + \left( C + \frac{2\pi^2}{3} B \right) (d + L)$$

$$= O \left( C \sqrt{(d+L)} O_{\text{max}} n \log(n) \right). \quad (15)$$

q.e.d.

C. Pseudocode of $idTSinc$

The pseudocode of $idTSinc$ is summarized in Algorithm 2.

Algorithm 2: A computational efficient variant of $idTSvi$

1: Input: $\epsilon > 0$
2: Randomly initialize $q$
3: for $t = 1,\ldots, n$ do
4: Sample $\theta_t$ proportionally to $q(\theta_t)$
5: Take action $a_t = \arg\max_{a} A_K r(a, \theta_t)$
6: Observe $x_t$ and receive reward $r(x_t, z_t)$
7: Randomly initialize $q$
8: Calculate $L(q)$ using (3) and set $L'(q) = -\infty$
9: while $L(q) - L'(q) \geq \epsilon$ do
10: Set $L'(q) = L(q)$
11: Update $q_t(z_t)$ using (4), for all $z_t$
12: Update $q(\theta)$ using (5)
13: Update $L(q)$ using (3)
14: end while
15: end for