Supplementary Material: Cheap Bandits

1. Proof of Proposition 1

For a given policy $\pi, \boldsymbol{\alpha}^*, T$, and a graph G define expected cumulative reward as

$$Regret(T, \pi, \boldsymbol{\alpha}^*, G) = \mathbb{E}\left[\sum_{t=1}^{T} \tilde{\boldsymbol{s}}_* \boldsymbol{\alpha}_* - \tilde{\boldsymbol{s}}_t \boldsymbol{\alpha}^* \Big| \boldsymbol{\alpha}^*\right]$$

where $\tilde{\boldsymbol{s}}_t = \pi'(t)\boldsymbol{Q}$, and $\boldsymbol{Q}$ is the orthonormal basis matrix corresponding to Laplacian of G. Let $\mathcal{G}_d$ denote the family of graphs with effective dimension d. Define T-period risk of the policy π

$$Risk(T, \pi) = \max_{G \in \mathcal{G}_d} \max_{\substack{\boldsymbol{\alpha}^* \in \mathcal{R}^N \\ \|\boldsymbol{\alpha}^*\|_{\boldsymbol{\Lambda}} < c}} \left[Regret\left(T, \pi, \boldsymbol{\alpha}^*, G\right)\right]$$

We first establish that there exists a graph with effective dimension d and a class of smooth reward functions defined over it with parameters $\boldsymbol{\alpha}^*$'s in a d-dimensional vector space.

Lemma 1. *Given T, there exists a graph $\widehat{G} \in \mathcal{G}_d$ such that*

$$\max_{\substack{\boldsymbol{\alpha}^* \in \mathcal{R}^d \\ \|\boldsymbol{\alpha}^*\|_{\boldsymbol{\Lambda}} < c}} \left[Regret\left(T, \pi, \boldsymbol{\alpha}^*, \widehat{G}\right)\right] \leq Risk(T, \pi)$$

Proof: We prove the lemma by the explicit construction of a graph. Consider a graph G consisting of d disjoint connected subgraphs denoted as $G_j : j = 1, 2 \ldots, d$. Let the nodes in each subgraph have the same reward. The eigenvalues of the graph are $\{0, \widehat{\lambda}_1, \cdots, \widehat{\lambda}_{N-d}\}$, where eigenvalue 0 is repeated d times. Note that the set of eigenvalues of the graph is the union of the set of eigenvalues of the individual subgraphs. Without loss of generality, assume that $\widehat{\lambda_1} > T/d \log(T/\lambda + 1)$. This is always possible, for example if subgraphs are cliques, which is what we assume. Then the effective dimension of the graph G is d. Since the graph separates into d disjoint subgraphs, we can split the reward function $\boldsymbol{f}_\alpha = \boldsymbol{Q}\boldsymbol{\alpha}$ into d parts, one corresponding to each subgraph. We write $\boldsymbol{f}_j = \boldsymbol{Q_j}\boldsymbol{\alpha}_j$ for $j = 1, 2, \ldots, d$, where $\boldsymbol{f}_j$ is the reward function associated with G_j, $\boldsymbol{Q}_j$ is the orthonormal matrix corresponding to Laplacian of G_j, and $\boldsymbol{\alpha}_i$ is a sub-vector of $\boldsymbol{\alpha}$ corresponding to node rewards on G_j.

Write $\boldsymbol{\alpha}_j = \boldsymbol{Q}_j' \boldsymbol{f}_j$. Since $\boldsymbol{f}_j$ is a constant vector, and except for one, all the columns in $\boldsymbol{Q}_j$ are orthogonal to $\boldsymbol{f}_j$, it is clear that $\boldsymbol{\alpha}_j$ has only one non-zero component. We conclude that for the reward functions that is constant on each subgraphs $\boldsymbol{\alpha}$ has only d non-zero components and is in a d-dimensional space. The proof of the lemma is completed by setting $\widehat{G} = G$. Note that a graph with effective dimension d cannot have more than d disjoint connected subgraphs. Next, we restrict our attention to graph $\widehat{G}$ and rewards that are piecewise constant on each clique. That means that the nodes in each clique have the same reward. Recall that action set $\mathcal{S}_D$ consists of actions that can probe a node or a group of neighboring nodes. Therefore, any group action will only allow us to observe average reward from a group of nodes within a clique but not across the cliques. Then, all node and group actions used to observe reward from within a clique are indistinguishable. Hence, the $\mathcal{S}_D$ collapses to set of d distinct actions one associated with each clique, and the problem reduces to that of selecting a clique with the highest reward. We henceforth treat each clique as an arm where all the nodes within the same clique share the same reward value.

We now provide a lower bound on the expected regret defined as follows

$$\widetilde{Risk}(T, \pi, \widehat{G}) = \mathbb{E}\left[Regret\left(T, \pi, \boldsymbol{\alpha}^*, \widehat{G}\right)\right], \tag{1}$$

where expectation is over the reward function on the arms.

To lower bound the regret we follow the argument of Auer et al. (2002) and their Theorem 5.1, where an adversarial setting is considered and the expectation in (1) is over the reward functions generated randomly according to Bernoulli distributions. We generalize this construction to our case with Gaussian noise. The reward generation process is as follows:

Without loss of generality choose cluster 1 to be the good cluster. At each time step t, sample reward of cluster 1 from the Gaussian distribution with mean $\frac{1}{2}+\xi$ and unit variance. For all other clusters, sample reward from the Gaussian distribution with mean $\frac{1}{2}$ and unit variance.

The rest of the proof of the arguments follows exactly as in the proof of Theorem 5.1 (Auer et al., 2002) except at their Equation 29. To obtain an equivalent version for Gaussian rewards, we use the relationship between the L_1 distance of Gaussian distributions and their KL divergence. We then apply the formula for the KL divergence between the Gaussian random variables to obtain equivalent version of their Equation 30. Now note that, $\log(1-\xi^2) \sim -\xi^2$ (within a constant). Then the proof follows silmilarly by setting $\xi = \sqrt{d/T}$ and noting that the L_2 norm of the mean rewards is bounded by c for an appropriate choice of λ.

2. Proof of Proposition 2

In the following, we first give some definitions and related results.

Definition 1 (k-way expansion constant by Lee et al., 2012)**.** *Consider a graph G and $\mathcal{X} \subset \mathcal{V}$. Let*

$$\phi_G(\mathcal{X}) := \phi(\mathcal{X}) = \frac{|\partial \mathcal{X}|}{V(\mathcal{X})},$$

where $V(\mathcal{X})$ denotes the sum of the degree of nodes in $\mathcal{X}$ and $|\partial\mathcal{X}|$ denotes the number of edges between the nodes in $\mathcal{X}$ and $\mathcal{V}\backslash\mathcal{X}$. For all $k>0$, k-way expansion constant is defined as

$$\rho_{\mathcal{G}}(k) = \min\left\{\max \phi(\mathcal{V}^i) : \cap_{i=1}^k \mathcal{V}^i = \varnothing, |\mathcal{V}^i| \neq 0\right\}.$$

Let $\mu_1 \leq \mu_2, \ldots, \leq \mu_N$ denote the eigenvalues of the normalized Laplacian of G.

Theorem 1 (Gharan & Trevisan (2014), Lee et al. (2012))**.** *Let $\varepsilon > 0$ and $\rho(k+1) > (1+\varepsilon)\rho(k)$ holds for some $k>0$. Then the following holds:*

$$\mu_k/2 \leq \rho(k) \leq \mathcal{O}(k^2)\sqrt{\mu_k} \tag{2}$$

There exist k partitions $\{\mathcal{V}^i : i = 1, 2, \ldots, k\}$ of $\mathcal{V}$ such that $\forall\ i = 1, 2, \ldots, k$

$$\phi(\mathcal{V}^i) \leq k\rho(k) \quad \textit{and} \tag{3}$$

$$\phi(G[\mathcal{V}^i]) \geq \varepsilon\rho(k+1)/14k \tag{4}$$

where $\phi(G[\mathcal{X}])$ denotes the Cheeger's constant (conduntance) of the subgraph induced by $\mathcal{X}$.

Definition 2 (Isoperimetric number)**.**

$$\theta(G) = \left\{\min \frac{\partial \mathcal{X}}{|\mathcal{X}|} : |\mathcal{X}| \leq \mathcal{X}/2\right\}.$$

Let $\lambda_1 \leq \lambda_2, \ldots, \leq \lambda_N$ denote the eigenvalues of the unnormalized Laplacian of G. We remind the reader of the following standard result.

$$\lambda_2/2 \leq \theta(G) \leq \sqrt{2\kappa\lambda_2}. \tag{5}$$

Proof: The relation $\lambda_{k+1}/\lambda_k \geq \mathcal{O}(k^2)$ implies that $\mu_{k+1}/\mu_k \geq \mathcal{O}(k^2)$. Using the upper and lower bounds on the eigenvalues in (2), the relation $\rho_{k+1} \geq (1+\varepsilon)\rho_k$ holds for some $\varepsilon > 1/2$. Then, applying Theorem 1 we get k-partitions satisfying (3)-(4). Let $\boldsymbol{L}_i$ denote the Laplacian induced by the subgraph $G[\mathcal{V}^j] = (\mathcal{V}^j, \mathcal{E}^j)$ for $j = 1, 2, \cdots k$. By the quadratic property of the graph Laplacian we have

$$\begin{aligned}\boldsymbol{f'Lf} &= \sum_{(u,v)\in\mathcal{E}} (f_u - f_v)^2 \\ &= \sum_{j=1}^{k} \sum_{(u,v)\in\mathcal{E}_j} (f_u - f_v)^2 \\ &= \sum_{j=1}^{k} \boldsymbol{f_j' L_j f_j}\end{aligned}$$

where $\boldsymbol{f}_j$ denotes the reward vector on the induced subgraph $G_j := G[\mathcal{V}^j]$. In the following we just focus on the optimal node. The same arguments holds for any other node. Without loss of generality assume that the node with optimal reward lies in subgraph G_l for some $1 \le l \le d$. From the last relation above we have $\boldsymbol{f}_l' \boldsymbol{l}_l \boldsymbol{f}_l \le c$. The reward functions on the subgraph G_l can be represented as $\boldsymbol{f}_l = \boldsymbol{Q}_l \boldsymbol{\alpha}_l$ for some $\boldsymbol{\alpha}_l$, where $\boldsymbol{Q}_l$ satisfies $\boldsymbol{L}_i = \boldsymbol{Q}_l' \boldsymbol{\Lambda}_L \boldsymbol{Q}_l$ and $\boldsymbol{\Lambda}_l$ denotes the diagonal matrix with eigenvalues of $\boldsymbol{\Lambda}_l$. We have

$$\begin{aligned}|F_G(\boldsymbol{s}_*) - F_G((\boldsymbol{s}_*^w)| &= |F_{G_l}(\boldsymbol{s}_*) - F_{G_l}(\boldsymbol{s}_*^w)| \\ &\le \|\boldsymbol{s}_* - \boldsymbol{s}_*^w\| \|\boldsymbol{Q}_l \boldsymbol{\alpha}_l\| \\ &\le \left(1 - \frac{1}{w}\right) \|\boldsymbol{Q}_l \boldsymbol{\Lambda}_l^{-1/2}\| \|\boldsymbol{\Lambda}_l^{1/2} \boldsymbol{\alpha}_l\| \\ &\le \frac{c}{\sqrt{\lambda_2(G_l)}} \quad \text{by Chauchy-Schwarz} \\ &\le \frac{\sqrt{2\kappa} c}{\theta(G_l)} \quad \text{from (5)} \\ &\le \frac{\sqrt{2\kappa} c}{\phi(G_l)} \quad \text{using } \theta(G_l) \ge \phi(G_l) \\ &\le \frac{14k\sqrt{2\kappa} c}{\varepsilon\rho(k+1)} \quad \text{from Theorem 1, Equation 4} \\ &\le \frac{56k\sqrt{2\kappa} c}{\mu_{k+1}} \quad \text{from Theorem 1, Equation 2} \\ &\le \frac{56k\kappa\sqrt{2\kappa} c}{\lambda_{k+1}} \quad \text{using } \mu_{k+1} \ge \lambda_{k+1}/\kappa.\end{aligned}$$

This completes the proof.

3. Analysis of CheapUCB

For a given confidence parameter δ define

$$\beta = 2R\sqrt{d \log\left(1 + \frac{T}{\lambda}\right) + 2\log\frac{1}{\delta}} + c,$$

and consider the ellipsoid around the estimate $\widehat{\boldsymbol{\alpha}}_t$

$$C_t = \{\boldsymbol{\alpha} : \|\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}\|_{V_t} \le \beta\}.$$

We first state the following results by Abbasi-Yadkori et al. (2011), Dani et al. (2008), and Valko et al. (2014) that we use later in our analysis.

Lemma 2 (self-normalized bound). *Let $\boldsymbol{\xi}_t = \sum_{i=1}^{t} \tilde{\mathbf{s}}_i \varepsilon_i$ and $\lambda > 0$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ and for all $t > 0$,*

$$\|\boldsymbol{\xi_t}\|_{\boldsymbol{V_t^{-1}}} \le \beta.$$

Lemma 3. *Let $V_0 = \lambda I$. We have:*

$$\log \frac{\det(V_t)}{\det(\lambda I)} \quad \leq \quad \sum_{i=1}^{t} \|\tilde{\mathbf{s}}_{\boldsymbol{i}}\|_{\mathbf{V_{i-1}^{-1}}} \|\tilde{\mathbf{s}}_i\|_{V_{i-1}^{-1}} \leq 2 \log \frac{\det(V_{t+1})}{\det(\lambda I)}$$

.

Lemma 4. *Let $\|\boldsymbol{\alpha}^*\|_2 \leq c$. Then, with probability at least $1-\delta$, for all $t \geq 0$ and for any $\mathbf{x} \in \mathcal{R}^n$ we have $\boldsymbol{\alpha}^* \in C_t$ and*

$$|\mathbf{x} \cdot (\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}^*)| \leq \|\mathbf{x}\|_{\mathbf{V_t^{-1}}} \beta.$$

Lemma 5. *Let d be the effective dimension and T be the time horizon of the algorithm. Then,*

$$\log \frac{\det(V_{T+1})}{\det(\Lambda)} \leq 2d \log \left(1 + \frac{T}{\lambda}\right).$$

3.1. Proof of Theorem 2

We first prove the case where degree of each node is at least $\log T$. Consider step $t \in [2^{j-1}, 2^j - 1]$ in stage $j = 1, 2, \cdots J-1$. Recall that in this step a probe of width $J-j+1$ is selected. Write $w_j := J-j+1$, and denote the probe of width $J-j+1$ associated with the optimal probe $\mathbf{s}_*$ as simply $\mathbf{s}_*^{w_j}$ and the corresponding GFT as $\tilde{\mathbf{s}}_*^{w_j}$. The probe selected at time t is denoted as $\mathbf{s}_t$. Note that both $\mathbf{s}_t$ and $\mathbf{s}_*^{w_j}$ lie in the set $\mathcal{S}_{J-j+1}$. For notational convenience let us denote

$$h(j) := \begin{cases} c'\sqrt{T}(J-j+1)/\lambda_{d+1} & \text{when (10) holds} \\ c'd/\lambda_{d+1} & \text{when (9) holds.} \end{cases}$$

The instantaneous regret in step t is

$$\begin{aligned}
r_t &= \tilde{\mathbf{s}}_* \cdot \boldsymbol{\alpha}^* - \tilde{\mathbf{s}}_t \cdot \boldsymbol{\alpha}^* \\
&\leq \tilde{\mathbf{s}}_*^{w_j} \cdot \boldsymbol{\alpha}^* + h(j) - \tilde{\mathbf{s}}_t \cdot \boldsymbol{\alpha}^* \\
&= \tilde{\mathbf{s}}_*^{w_j} \cdot (\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}}_t) + \tilde{\mathbf{s}}_*^{j} \cdot \widehat{\boldsymbol{\alpha}}_t + \beta \|\tilde{\mathbf{s}}_*^{w_j}\|_{\mathbf{V_t^{-1}}} - \beta \|\tilde{\mathbf{s}}_*^{w_j}\|_{\mathbf{V_t^{-1}}} - \tilde{\mathbf{s}}_t \cdot \boldsymbol{\alpha}^* + h(j) \\
&\leq \tilde{\mathbf{s}}_*^{w_j} \cdot (\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}}_t) + \tilde{\mathbf{s}}_t \cdot \widehat{\boldsymbol{\alpha}}_t + \beta \|\tilde{\mathbf{s}}_t\|_{\mathbf{V_t^{-1}}} - \beta \|\tilde{\mathbf{s}}_*^{w_j}\|_{\mathbf{V_t^{-1}}} - \tilde{\mathbf{s}}_t \cdot \boldsymbol{\alpha}^* + h(j) \\
&= \tilde{\mathbf{s}}_*^{w_j} \cdot (\boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}}_t) + \tilde{\mathbf{s}}_t \cdot (\widehat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}^*) + \beta \|\tilde{\mathbf{s}}_t\|_{V_t^{-1}} - \beta \|\tilde{\mathbf{s}}_*^{w_j}\|_{\mathbf{V_t^{-1}}} + h(j) \\
&\leq \beta \|\tilde{\mathbf{s}}_*^{w_j}\|_{\mathbf{V_t^{-1}}} + \beta \|\tilde{\mathbf{s}}_t\|_{\mathbf{V_t^{-1}}} + \beta \|\tilde{\mathbf{s}}_t\|_{\mathbf{V_t^{-1}}} - \beta \|\tilde{\mathbf{s}}_*^{w_j}\|_{\mathbf{V_t^{-1}}} + h(j) \\
&= 2\beta \|\tilde{\mathbf{s}}_t\|_{\mathbf{V_t^{-1}}} + h(j).
\end{aligned}$$

We used (9)/(10) in the first inequality. The second inequality follows from the algorithm design and the third inequality follows from Lemma 4. Now, the cumulative regret of the algorithm is given by

$$\begin{aligned}
R_T \leq & \sum_{j=1}^{J} \sum_{t=2^{j-1}}^{2^j-1} \min \left\{2, 2\beta \|\tilde{\mathbf{s}}_t\|_{V_t^{-1}} + h(j)\right\} \\
\leq & \sum_{j=1}^{J} \sum_{t=2^{j-1}}^{2^j-1} \min \left\{2, 2\beta_t \|\tilde{\mathbf{s}}_t\|_{V_t^{-1}}\right\} + \sum_{j=1}^{J-1} \sum_{t=2^{j-1}}^{2^j-1} h(j) \\
\leq & \sum_{t=1}^{T} \min \left\{2, 2\beta_t \|\tilde{\mathbf{s}}_t\|_{V_t^{-1}}\right\} + \sum_{j=1}^{J-1} h(j) 2^{j-1}.
\end{aligned}$$

Note that the summation in the second term includes only the first $J-1$ stages. In the last stage J, we use probes of width 1 and hence we do not need to use (9) or (10) in bounding the instantaneous regret. Next, we bound each term in the regret separately.

To bound the first term we use the same steps as in the proof of Theorem 1 of Valko et al. (2014). We repeat the steps below for convenience.

$$\begin{aligned}
\sum_{t=1}^{T} \min\{2, 2\beta\|\tilde{\mathbf{s}}_t\|_{V_t^{-1}}\} &\leq (2+2\beta)\sum_{t=1}^{T} \min\left\{1, \|\tilde{\mathbf{s}}_t\|_{\mathbf{V_t^{-1}}}\right\} \\
&\leq (2+2\beta)\sqrt{T\sum_{t=1}^{T} \min\left\{1, \beta_t\|\tilde{\mathbf{s}}_t\|_{\mathbf{V_t^{-1}}}\right\}^2} \\
&\leq 2(1+\beta)\sqrt{2T\log(|\boldsymbol{V}_{T+1}|/|\boldsymbol{\Lambda}|)} \qquad (6) \\
&\leq 4(1+\beta)\sqrt{Td\log(1+T/\lambda)} \qquad (7) \\
&\leq \left(8R\sqrt{2\log\frac{1}{\delta} + d\log\left(1+\frac{T}{\lambda}\right)} + 4c + 4\right)\sqrt{Td\log\left(1+\frac{T}{\lambda}\right)}.
\end{aligned}$$

We used Lemma 3 and 5 in inequalities (6) and (7) respectively. The final bound follows from plugging in the value of β.

3.2. For the case when (10) holds:

For this case we use $h(j) = c'\sqrt{T}(J-j+1)/\lambda_{d+1}$. First observe that $2^{j-1}h(j)$ is increasing in $1 \leq j \leq J-1$. We have

$$\begin{aligned}
\sum_{j=1}^{J-1} \frac{2^{j-1}c'\sqrt{T}(J-j+1)}{\lambda_{d+1}} &\leq (J-1)\frac{2^{J-1}\sqrt{T}c'}{\lambda_{d+1}} \\
&\leq (J-1)\frac{2^{\log_2 T-1}c'\sqrt{T}}{\lambda_{d+1}} \\
&\leq (J-1)\frac{c'\sqrt{T}(T/2)}{(T/d\log(T/\lambda+1))} \\
&\leq dc'\sqrt{T/4}\log_2(T/2)\log(T/\lambda+1).
\end{aligned}$$

In the second line we applied the definition of effective dimension.

3.3. For the case when $\lambda_{d+1}/\lambda_d \geq \mathcal{O}(d^2)$

For the case $\lambda_{d+1}/\lambda_d \geq \mathcal{O}(d^2)$ we use $h(j) = c'd/\lambda_{d+1}$.

$$\begin{aligned}
\sum_{j=1}^{J-1} \frac{2^{j-1}c'd}{\lambda_{d+1}} &\leq \frac{2^{J-1}c'd}{\lambda_{d+1}} \\
&\leq \quad c'd^2\log_2(T/2)\log(T/\lambda+1).
\end{aligned}$$

Now consider the case where minimum degree of the nodes is $1 < a \leq \log T$. In this case, we modify the algorithm to use only signals of width a in the first $\log T - a + 1$ stages and subsequently the signal width is reduced by one in each of the following stages. The previous analysis holds for this case and we get the same bounds on the cumulative regret and cost. When $a = 1$, CheapUCB is same as the SpectralUCB, hence total cost and regret is same as that of SpectralUCB.

To bound the total cost, note that in stage j we use signals of width $J-j+1$. Also, the cost of a signal given

in (2) can be upper bounded as $C(\mathbf{s}_i^w) \leq \frac{1}{w}$. Then, we can upperbound total cost of signals used till step T as

$$\begin{aligned}
&\sum_{j=1}^{J} \frac{2^{j-1}}{J-j+1} \\
&\leq \frac{1}{2}\sum_{j=1}^{J-1} 2^{j-1} + \frac{T}{2} \\
&\leq \frac{1}{2}\left(\frac{T}{2} - 1\right) + \frac{T}{2} \\
&= \frac{3T}{4} - \frac{1}{2}.
\end{aligned}$$

References

Abbasi-Yadkori, Yasin, Pál, David, and Szepesvári, Csaba. Improved Algorithms for Linear Stochastic Bandits. In *Neural Information Processing Systems*. 2011.

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav, and Schapire, Robert E. The Nonstochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, 32(1):48–77, January 2002.

Dani, Varsha, Hayes, Thomas P, and Kakade, Sham M. Stochastic Linear Optimization under Bandit Feedback. In *Conference on Learning Theory*, 2008.

Gharan, S. O. and Trevisan, L. Partitioning into expanders. In *Proceeding of Symposium of Discrete Algorithms, SODA*, Portland, Oregon, USA, 2014.

Lee, James R., Gharan, Shayan Oveis, and Trevisan, Luca. Multi-way spectral partitioning and higher-order cheeger inequalities. In *Proceeding of STOC*, 2012.

Valko, Michal, Munos, Rémi, Kveton, Branislav, and Kocák, Tomáš. Spectral Bandits for Smooth Graph Functions. In *31th International Conference on Machine Learning*, 2014.