A. Causal Graphs and Interventions

A causal graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ specifies causal relationships among the random variables representing the vertices of the graph $\mathcal{V}$. The relationships are specified by the directed edges $\mathcal{E}$; an edge $V_i \to V_j$ implies that $V_i \in \mathcal{V}$ is a direct parental cause for the effect $V_j \in \mathcal{V}$. With some abuse of notation, we will denote the random variable associated with a node $V \in \mathcal{V}$ by V itself. We will denote the parents of a node V by $pa(V)$. The causal dependence implies that $V = f_V(U \in pa(V), \epsilon_V)$, where ϵ_V is an independent exogenous noise variable. One does not get to measure the functions f_V in practice. The noise variable and the above functional dependence induce a conditional probability distribution $\mathrm{P}(V|pa(V))$. Further, the joint distribution of $\{V\}_{V\in\mathcal{V}}$ decomposes into product of conditional distributions according to $\mathcal{G}$ viewed as a Bayesian Network, i.e. $\mathrm{P}(\{V\}_{V\in\mathcal{V}}) = \prod_{V\in\mathcal{V}} \mathrm{P}(V|pa(V))$.

Interventions in a causal setting can be categorized into two kinds:

1. *Soft Interventions*: At node V, the conditional distribution relating $pa(V)$ and V is changed to $\tilde{\mathrm{P}}(V|pa(V))$.
2. *Hard Interventions:* We force the node V to take a specific value x. The conditional distribution $\tilde{\mathrm{P}}(V|pa(V))$ is set to a point mass function $\mathbf{1}_{V=x}$.

B. Variations of the Problem Setting

In this section we provide more general causal settings where our results can be directly applied.

Multiple nodes at the graph: This is illustrated in Fig. 5a. Soft interventions can be performed at multiple nodes like at $\mathcal{V} = \{V_1, V_2\}$. These interventions can be modeled as changing the distribution $P(\mathcal{V}|pa(\mathcal{V}))$ where $pa(\mathcal{V})$ are the union of parents of V_1 and V_2. These distributions can be thought of as the arms of the bandits and our techniques can be applied as before to estimate the best intervention.

Directed cut between sources and targets: Fig. 5b represents the most general scenario in which our techniques can be applied. Soft or hard interventions can be performed at multiple *source* nodes, while the goal is to choose the best out of these interventions in terms of maximizing a known function of multiple target nodes. If the effect of these interventions can be estimated on a directed cut separating the targets and the sources then our techniques can be applied as before. This is akin to knowing $P(V_1, V_2)$ under all the interventions in Fig. 5b, because V_1 and V_2 is a directed cut separating the sources and the targets.

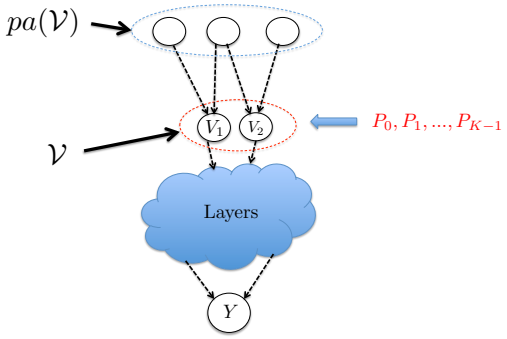
Empirical knowledge of continuous arm distributions: Our techniques can be applied to continuous distributions $\mathrm{P}(V|pa(V))$ as shown in our empirical results in Section 4.1. The extension is straight-forward by using the general definition of f-divergences. More importantly our techniques can be applied even if only prior empirical samples from the distributions $\mathrm{P}(V|pa(V))$ is available and not the whole distributions. In this case the f-divergences can be estimated using nearest neighbor estimators similar to (Pérez-Cruz, 2008). Moreover, for the importance sampling only ratios of distributions are needed, which can again be estimated using nearest neighbor based techniques from empirical data.

C. Discussion and Future Work

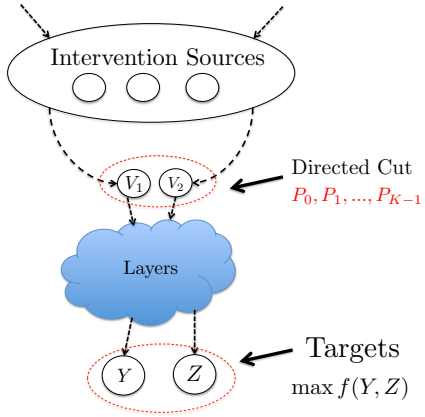
In this paper, we analyze the problem of identifying the best arm at a node V in a causal graph (various known conditionals $\mathrm{P}_k(V|pa(V))$) in terms of its effect on a *target* variable Y further downstream, possibly in a less understood portion of the larger causal network. We characterize the hardness of this problem in terms of the relative divergences of the various conditionals that are being tested and the gaps between the expected value of the target under the various arms. We provide the first problem dependent simple regret and error bounds for this problem, that is a natural generalization of (Audibert & Bubeck, 2010), but with information leakage between arms. We provide an efficient successive rejects style algorithm that achieves these guarantees, by leveraging the leakage of information, through carefully designed clipped importance samplers. Further, we introduce a new f-divergence measure that may be relevant for analyzing importance sampling estimators in the causal context. This may be of independent interest. We believe that our work paves the way for various interesting problems with significant practical implications. In the following, we state a few open questions in this regard:

Tighter guarantees on SRISv2: In Section 4, we have observed that a slightly modified version of our algorithm SRISv2 performs the best among all the competing algorithms including SRISv1. The only difference of SRISv2 from Algorithm 1, is that in line 6 the estimators used in a phase also uses samples from past phases, but clipped according to the criterion in the current phase. We believe that this algorithm has tighter error and simple regret guarantees. We conjecture that at least one of the $\log(1/\Delta_k)$, in the definition of $\bar{H}$ in (4) can be eliminated, thus leading to better guarantees.

Estimating the marginals of the parents: In Algorithm 1, either the marginals of the parents of V, that is $\mathrm{P}(pa(V))$ is required in order to calculate the f-divergences in Definition 2, or prior data involving the parents is required to estimate the f-divergences directly from data. However, we believe it is possible to model this estimation, directly into the online framework, as data about the marginals of the



(a) Illustration of a scenario where there are multiple intervention sources $\mathcal{V} = \{V_1, V_2\}$. Each soft intervention is a change in the distribution $\mathrm{P}(\mathcal{V}|pa(\mathcal{V}))$.



(b) Illustration of a causal setting, where there are many intervention sources. Soft or hard interventions can be performed at these sources and the effects of these interventions can be observed at the target node. Our techniques can be applied in choosing the best intervention in terms of maximizing a function on the target nodes, provided the effects of these interventions on $P(V_1, V_2)$ are known. Here V_1 and V_2 form a directed cut separating the sources and the targets.

Figure 5. Illustration of more general causal settings where our algorithms can be directly applied.

parents are available through the samples in all the arms, as these marginals remain unchanged.

Problem Dependent Lower Bound: In (Lattimore et al., 2016), a problem independent lower bound of $O(1/\sqrt{T})$ has been provided for a special causal graph. However, the problem parameter dependent lower bound like that of (Audibert & Bubeck, 2010) still remains an open problem. We believe that the lower bound will depend on the divergences between the distributions and the gaps between the rewards of the arms, similar to the term in (4).

General Learning Framework: Our work paves the way for a more general setting for learning counterfactual effects. Importance sampling is a fairly general tool and can be ideally applied at any set of nodes of a causal graph. So, in principle it is possible to study the effect of a change at V on a target Y, by using importance sampling between the changed marginal distributions at an intermediate cut $\mathcal{S}$ that blocks every path from V to Y. In fact, this is explored in a non-bandit context in (Bottou et al., 2013). An important question is: What is the most suitable cut to be used? (Lattimore et al., 2016) uses the cut closest to Y, i.e. immediate parents of Y. However, the marginals of the cut under different changes need to be estimated this 'far' from the source closer to the target. Therefore, there is a trade-off that involves a delicate balance between the estimation errors of the changes at an intermediate cut between V and Y, and the reduction in importance sampling divergences between cut distributions closer to the target Y. We believe understanding this is quite important to fully exploit partial/full knowledge about causal graph structure to answer causal strength questions from data observed.

D. Interpretation of our Theoretical Results

In this section we compare our theoretical bounds on the probability of mis-identification with the corresponding bounds in (Audibert & Bubeck, 2010). We also compare our simple regret guarantees with the guarantees in (Lattimore et al., 2016). In both these cases, we demonstrate significant improvements. These theoretical improvements are exhibited in our empirical results in Section 4.1.

D.1. Comparison with (Audibert & Bubeck, 2010)

Let $\tilde{\mathcal{R}}(\Delta_k) = \{s : \Delta_s \leq \Delta_k\}$, i.e. the set of arms which are closer to the optimal than arm k. Let $\tilde{H} = \max_{k \neq k*} \frac{|\tilde{\mathcal{R}}(\Delta_k)|}{\Delta_k^2}$. The result for the best arm identification with no information leakage in (Audibert & Bubeck, 2010) can be stated as: *The error in finding the optimal arm is bounded as*:

$$e(T) \leq O\left(K^2 \exp\left(-\frac{T-K}{\overline{\log}(K)\tilde{H}}\right)\right) \tag{6}$$

One intuitive interpretation for $\tilde{H}$ is that it is the maximum among the number of samples (neglecting some log factors) required to conclude that arm k is suboptimal from

among the arms which are closer to the optimal than itself. Intuitively, this is because when there is no information leakage, one requires $1/\Delta_k^2$ samples to distinguish between the k-th optimal arm and the optimal arm. Further, the kth arm is played only $1/k$ fraction of the times since we do not know the identity of the k-th optimal arm.

Our main result in Theorem 1 can be seen to be a generalization of the existing result for the case when there is information leakage between the arms (various changes in a causal graph).

The term $\sigma^*(B, \mathcal{R}^*(\Delta_k))$ in our setting is the 'effective standard deviation' due to information leakage. There is a similar interpretation of our result (ignoring the log factors): Since there is information leakage, the expression $\frac{(\sigma^*)^2}{(\Delta_k)^2}$ characterizes the number of samples required to weed out arm k out of contention from among competing arms (arms that are at a distance at most twice than that of arm k from the optimal arm). The interpretation of 'effective variance' is justified using importance sampling which is detailed in Section E.3. Further, in our framework σ^* also incorporates any budget constraint that comes with the problem, i.e. any *apriori* constraint on the relative fraction of times different arms need to be pulled.

For ease of exposition let (k) denote the index of the k-th best arm (for $k = 1, .., K$) and $\Delta_{(k)}$ denotes the corresponding gap. In this setting, the terms $\tilde{H}$ (from the result in (Audibert & Bubeck, 2010)) and $\bar{H}$ can be written as:

$$\tilde{H} = \max_{k \neq 1} \frac{k}{\Delta_{(k)}^2}$$
$$\bar{H} = \max_{k \neq 1} \frac{\sigma^*(B, \mathcal{R}^*(\Delta_{(k)}))^2}{\Delta_{(k)}^2}.$$

$\sigma^*(B, \mathcal{R}^*(\Delta_{(k)}))$ can be smaller than $\sqrt{k}$ due to information leakage as every single arm pull contributes to another arm's estimate. Therefore, these provide better guarantees than (Audibert & Bubeck, 2010).

To see the improvement over the previous result in (Audibert & Bubeck, 2010), we consider a special case when the cost budget B is infinity and there is only the the sample budget T. In addition, let us assume that the log divergences are such that: $M_{ij} \leq \eta M_{ii} = \eta << \sqrt{|\mathcal{R}|}$, $\forall i \neq j$. Let $\mathcal{R} = \mathcal{R}^*(\Delta_{(k)})$. If $\eta > \sqrt{|\mathcal{R}|}$, the optimal solution for (2) is a bit complicated to interpret. Consider the feasible allocation $\nu_i = \frac{1}{|\mathcal{R}|}$, $\forall i \in \mathcal{R}$ in (2). Evaluating the objective function for this feasible allocation, it is possible to show that $\sigma^* \leq \frac{\eta}{1-1/|\mathcal{R}|} << \sqrt{|\mathcal{R}|}$. Hence, unless the variance due to information leakage is too bad, the effective variance is smaller than that of the case with no information leakage.

The improvement over the no information leakage setting, is even more pronounced under budget constraints. Consider the setting **S1**, and assume that the fractional budget of the *difficult* arms, $B = o(1)$. This implies that the total number of samples available for difficult arms is $o(T)$. The budget constrained case has not been analyzed in (Audibert & Bubeck, 2010), however in the absence of information leakage, one would expect that the arms with the least number of samples would be the most difficult to eliminate, and therefore the error guarantees would scale as $\exp(-O(BT)/\tilde{H}) \sim \exp(-o(T)/\tilde{H})$ (excluding log factors). On the other hand, our algorithm can leverage the information leakage and the error guarantee would scale as $\exp(-O(T)/\bar{H})$, which can be order-wise better if the *effective standard deviations* are well-behaved.

D.2. Comparison with (Lattimore et al., 2016)

In (Lattimore et al., 2016), the algorithm is based on clipped importance samples, where the clipper is always set at a static level of $O(\sqrt{T})$ (excluding log factors). The simple regret guarantee in (Lattimore et al., 2016) scales as $O(\sqrt{(m(\eta)/T)\log T})$, where $m(\eta)$ is a global hardness parameter. The guarantees do not adapt to the problem parameters, specifically the gaps $\{\Delta_k\}_{k\in[K]}$.

On the contrary, we provide problem dependent bounds, which differentiates the arms according to its gap from the optimal arm and its *effective standard deviation* parameter. The terms $\bar{H}_k$ can be interpreted as the hardness parameter for rejecting arm k. Note that $\bar{H}_k$ depends only on the arms that are at least as *bad* in terms of their gap from the optimal arm. Moreover the guarantees are adapted to our general budget constraints, which is absent in (Lattimore et al., 2016). It can be seen that when Δ_k's do not scale in T, then our simple regret is exponentially small in T (dependent on $\bar{H}_k$'s) and can be much less than $O(1/\sqrt{T})$. The guarantee also generalizes to the problem independent setting when Δ_k's scale as $O(1/\sqrt{T})$.

E. Proofs

In this section we present the theoretical analysis of our algorithm. Before we proceed to the proof of our main theorems, we derive some key lemmas that are useful in analyzing clipped importance sampled estimators.

E.1. Clipped Importance Sampling Estimator

One of the salient features of this problem is that there is information leakage among the K arms of the bandit. The different arms that are being tested only differ in the conditional distribution of V given its parents $pa(V)$, while the rest of the relationships in the causal graph $\mathcal{G}$ remain unchanged. Since the different candidate conditional distributions $\mathrm{P}_k(V|pa(V))$ are known from prior knowledge,

it is possible to utilize samples obtained under an arm j to obtain an estimate for the expectation under arm i (i.e $\mathbb{E}_i[Y]$). We will see in subsequent sections that the *goodness* of samples obtained under arm j for estimating $\mathbb{E}_i[Y]$, is dependent on a particular divergence metric between the distributions $\mathrm{P}_i(V|pa(V))$ and $\mathrm{P}_j(V|pa(V))$. A popular method for utilizing this information leakage among different distributions is *importance sampling*. Importance sampling has been used before in counterfactual analysis in a similar causal setting (Lattimore et al., 2016; Bottou et al., 2013). In the subsequent sections, we introduce importance samplers in the context of our problem and provide some novel techniques to analyze the confidence intervals for the importance samplers.

Importance Sampling: Now we introduce the concept of importance sampling which is one of the key tools we use to leverage the information leakage between the candidate arms. Suppose we get samples from arm $j \in [K]$ and we are interested in estimating $\mathbb{E}_i[Y]$. In this context it helpful to express $\mathbb{E}_i[Y]$ in the following manner:

$$\mathbb{E}_i[Y] = \mathbb{E}_j\left[Y\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))}\right] \tag{7}$$

(7) is trivially true because the only change to the joint distribution of all the variables in the causal graph $\mathcal{G}$ under arm i and j is at the factor $\mathrm{P}(V|pa(V))$. Suppose we observe t samples of $\{Y, V, pa(V)\}$ from the arm j, denoted by $\{Y_j(s), V_j(s), pa(V)_j(s)\}_{s=1}^t$. Under the observation of Equation (7), one might assume that the naive estimator,

$$\hat{Y}_i'(j) = \frac{1}{t}\sum_{s=1}^{t} Y_j(s)\frac{\mathrm{P}_i(V_j(s)|pa(V)_j(s))}{\mathrm{P}_j(V_j(s)|pa(V)_j(s))} \tag{8}$$

provides a good estimate for $\mu_i = \mathbb{E}_i[Y]$. However, the confidence guarantees on such an estimate can be arbitrarily bad as even though Y is bounded. This is because the factor $\mathrm{P}_i(V|pa(V))/\mathrm{P}_j(V|pa(V))$ can be very large for several values of $V, pa(V)$. Therefore, usual confidence inequalities like Azuma-Hoeffding's, Berstein's would not yield good confidence intervals.

Clipped Importance Samplers: In the previous section, we observe that the naive estimator of (8) is not suitable for yielding good confidence intervals. It has been observed in the context of importance sampling, that clipping the estimator in (8) at a carefully chosen value, can yield better confidence guarantees even though the resulting estimator will become slightly biased (Bottou et al., 2013). Before we introduce the precise estimator, let us define a key quantity that will be useful for the analysis.

Definition 4. *We define $\eta_{i,j}(\epsilon)$ as follows:*

$$\eta_{i,j}(\epsilon) = \left\{\operatorname*{argmin}_{\eta} : \mathbb{P}_i\left(\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))} > \eta\right) \leq \frac{\epsilon}{2}\right\} \tag{9}$$

for all $i, j \in [K]$, where $\epsilon > 0$.

We shall see that the $\eta_{i,j}(\epsilon)$ is related to the conditional f-divergence between $\mathrm{P}_i(V|pa(V))$ and $\mathrm{P}_j(V|pa(V))$ for the carefully chosen function $f_1(.)$ as introduced in Section 3.1.

Now we are at a position to provide confidence guarantees on the following clipped estimator:

$$\begin{aligned}&\hat{Y}_i^{(\eta)}(j)\\ &= \frac{1}{t}\sum_{s=1}^{t} Y_j(s)\frac{\mathrm{P}_i(V_j(s)|pa(V)_j(s))}{\mathrm{P}_j(V_j(s)|pa(V)_j(s))}\times\\ &\mathbb{1}\left\{\frac{\mathrm{P}_i(V_j(s)|pa(V)_j(s))}{\mathrm{P}_j(V_j(s)|pa(V)_j(s))} \leq \eta_{ij}(\epsilon)\right\}.\end{aligned} \tag{10}$$

Lemma 1. *The estimate $\hat{Y}_i^{(\eta)}(j)$ for $\eta = \eta_{i,j}(\epsilon)$ satisfies the following:*

1.

$$\mathbb{E}_j\left[\hat{Y}_i^{(\eta)}(j)\right] \leq \mu_i \leq \mathbb{E}_j\left[\hat{Y}_i^{(\eta)}(j)\right] + \frac{\epsilon}{2} \tag{11}$$

2.

$$\begin{aligned}&\mathbb{P}\left(\mu_i - \delta - \epsilon/2 \leq \hat{Y}_i^{(\eta)}(j) \leq \mu_i + \delta\right)\\ &\geq 1 - 2\exp\left(-\frac{\delta^2 t}{2\eta_{i,j}(\epsilon)^2}\right).\end{aligned} \tag{12}$$

Proof. We have the following chain:

$$\begin{aligned}&\mathbb{E}_j\left[Y\frac{\mathrm{P}_i(V|pa(V)}{\mathrm{P}_j(V|pa(V))}\right]\\ &= \mathbb{E}_j\left[Y\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))}\mathbb{1}\left\{\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))} \leq \eta_{i,j}(\epsilon)\right\}\right]\\ &+ \mathbb{E}_j\left[Y\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))}\mathbb{1}\left\{\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))} > \eta_{i,j}(\epsilon)\right\}\right]\\ &\overset{(a)}{\leq} \mathbb{E}_j\left[Y\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))}\mathbb{1}\left\{\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))} \leq \eta_{i,j}(\epsilon)\right\}\right]\\ &+ \mathbb{P}_i\left(\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))} > \eta_{i,j}(\epsilon)\right)\\ &\leq \mathbb{E}_j\left[Y\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))}\mathbb{1}\left\{\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))} \leq \eta_{i,j}(\epsilon)\right\}\right] + \frac{\epsilon}{2}\end{aligned}$$

Here, (a) is because $Y \in [0, 1]$. This yields the first part of the lemma:

$$\mathbb{E}_j\left[\hat{Y}_i^{(\eta)}(j)\right] \leq \mu_i \leq \mathbb{E}_j\left[\hat{Y}_i^{(\eta)}(j)\right] + \frac{\epsilon}{2} \tag{13}$$

where $\eta = \eta_{i,j}(\epsilon)$. Note that all the terms in the summation of (10) are bounded by $\eta_{i,j}(\epsilon)$. Therefore, by an application

of Azuma-Hoeffding we obtain:

$$\mathbb{P}\left(|\hat{Y}_i^{(\eta)}(j) - \mathbb{E}_j\left[\hat{Y}_i^{(\eta)}(j)\right]| > \delta\right) \leq 2\exp\left(-\frac{\delta^2 t}{2\eta_{i,j}(\epsilon)^2}\right) \tag{14}$$

Combining Equation (13) and (14), we obtain the first part of our lemma. □

E.2. Relating $\eta_{ij}(\cdot)$ with f-divergence

Now we are left with relating $\eta_{i,j}(\epsilon)$ to a particular f-divergence (D_{f_1} defined in Section 3.1) between $\mathrm{P}_i(V|pa(V))$ and $\mathrm{P}_j(V|pa(V))$. We have the following relation,

$$\mathbb{E}_i\left[\exp\left(\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))}\right)\right] = \left[1 + D_{f_1}\left(\mathrm{P}_i \| \mathrm{P}_j\right)\right] e. \tag{15}$$

The following lemma expresses the quantity $\eta_{i,j}(\epsilon)$ as a separable function of $D_{f_1}\left(\mathrm{P}_i \| \mathrm{P}_j\right)$ and ϵ, and is one of the key tools used in subsequent analysis.

Lemma 2. *It holds that,* $\eta_{i,j}(\epsilon) \leq \log\left(\frac{2}{\epsilon}\right) + 1 + \log\left(1 + D_{f_1}\left(\mathrm{P}_i \| \mathrm{P}_j\right)\right)$. *Furthermore,*

$$\eta_{i,j}(\epsilon) \leq 2\log\left(\frac{2}{\epsilon}\right)\left[1 + \log(1 + D_{f_1}(\mathrm{P}_i \| \mathrm{P}_j))\right] \tag{16}$$

when, $\epsilon \leq 1$.

Proof. We have the following chain:

$$\mathbb{P}_i\left(\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))} > \eta\right) \tag{17}$$

$$= \mathbb{P}_i\left(\exp\left(\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))}\right) > \exp(\eta)\right)$$

$$\overset{(a)}{\leq} \mathbb{E}_i\left[\exp\left(\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))}\right)\right]\exp(-\eta) \tag{18}$$

(a) - We used Markov's inequality. Suppose, we have the right hand side to be at most $\epsilon/2$. Then we have,

$$\mathbb{E}_i\left[\exp\left(\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))}\right)\right]\exp(-\eta) \leq \epsilon/2 \tag{19}$$

Now using (15), we have:

$$\eta \geq \log\left(\frac{2}{\epsilon}\right) + 1 + \log(1 + D_{f_1}(\mathrm{P}_i \| \mathrm{P}_j)) \tag{20}$$

$$\implies \mathbb{P}_i\left(\frac{\mathrm{P}_i(V|pa(V))}{\mathrm{P}_j(V|pa(V))} > \eta\right) \leq \epsilon/2$$

From, the definition of $\eta_{i,j}(\epsilon)$, we have:

$$\eta_{i,j}(\epsilon) \leq \log\left(\frac{2}{\epsilon}\right) + 1 + \log(1 + D_{f_1}(\mathrm{P}_i \| \mathrm{P}_j))$$

$$\overset{a}{\leq} 2\log\left(\frac{2}{\epsilon}\right)\left[1 + \log(1 + D_{f_1}(\mathrm{P}_i \| \mathrm{P}_j))\right], \ \forall \epsilon \leq 1 \tag{21}$$

(a) - This is due to the inequality $p + q \leq 2pq$ when $q \geq 1$ and $p \geq \log_e(2)$. □

Now, we introduce the main result of this section as Theorem 3. Recall that $M_{ij} = 1 + \log(1 + D_{f_1}(\mathrm{P}_i \| \mathrm{P}_j))$.

Theorem 3. *The estimate* $\hat{Y}_i^{(\eta)}(j)$ *for* $\eta = 2\log(2/\epsilon)M_{ij}$ *satisfies the following confidence guarantees:*

$$\mathbb{P}\left(\mu_i - \delta - \epsilon/2 \leq \hat{Y}_i^{(\eta)}(j) \leq \mu_i + \delta\right)$$

$$\geq 1 - 2\exp\left(-\frac{\delta^2 t}{8\log(2/\epsilon)^2 M_{ij}^2}\right).$$

Proof. The proof is immediate from Lemmas 2 and 1. □

E.3. Aggregating Heterogenous Clipped Estimators

In Section E.1, we have seen how samples from one of the candidate distribution can be used for estimating the target mean under another arm. Therefore, it is possible to obtain information about the target mean under the k^{th} arm ($\mathbb{E}_k[Y]$) from the samples of all the other arms. It is imperative to design an efficient estimator of $\mathbb{E}_k[Y]$ ($\forall k \in [K]$) that seamlessly uses the samples from all arms, possibly with variable weights depending on the relative divergences between the distributions. In this section we will come up with one such estimator, based on the insight gained in Section E.1.

Recall the quantities $M_{kj} = 1 + \log(1 + D_{f_1}(\mathrm{P}_k \| \mathrm{P}_j))$ ($\forall k, j \in [K]$). These quantities will be the key tools in designing the estimators in this section. Suppose we obtain τ_i samples from arm $i \in [K]$. Let the total number of samples from all arms put together be τ.

Let us index all the samples by $s \in \{1, 2, .., \tau\}$. Let $\mathcal{T}_k \subset \{1, 2, .., \tau\}$ be the indices of all the samples collected from arm k. Further, let $Z_k = \sum_{j\in[K]} \tau_j / M_{kj}$. Now, we are at the position to introduce the estimator for μ_k, which we will denoted by $\hat{Y}_k^{\epsilon}$ (ϵ is an indicator of the level of confidence desired):

$$\hat{Y}_k^{\epsilon} = \frac{1}{Z_k}\sum_{j=0}^{K}\sum_{s\in\mathcal{T}_j}\frac{1}{M_{kj}}Y_j(s)\frac{\mathrm{P}_k(V_j(s)|pa(V)_j(s))}{\mathrm{P}_j(V_j(s)|pa(V)_j(s))} \times$$

$$\mathbb{1}\left\{\frac{\mathrm{P}_k(V_j(s)|pa(V)_j(s))}{\mathrm{P}_j(V_j(s)|pa(V)_j(s))} \leq 2\log(2/\epsilon)M_{kj}\right\}. \tag{22}$$

In other words, $\hat{Y}_k^\epsilon$ is the weighted average of the clipped samples, where the samples from arm j are weighted by $1/M_{kj}$ and clipped at $2\log(2/\epsilon)M_{kj}$.

Lemma 3.

$$\hat{\mu}_k := \mathbb{E}\left[\hat{Y}_k^\epsilon\right] \leq \mu_k \leq \mathbb{E}\left[\hat{Y}_k^\epsilon\right] + \frac{\epsilon}{2} \tag{23}$$

Proof. We note that $\hat{Y}_k^\epsilon$ can be written as:

$$\hat{Y}_k^\epsilon = \frac{1}{Z_k}\sum_{j=0}^{K}\frac{\tau_j}{M_{kj}}\tilde{Y}_{kj}^\epsilon \tag{24}$$

Here, $\tilde{Y}_{kj}^\epsilon = \frac{1}{\tau_j}\sum_{s\in\mathcal{T}_j} Y_j(s)\frac{\mathrm{P}_k(V_j(s)|pa(V)_j(s))}{\mathrm{P}_j(V_j(s)|pa(V)_j(s))}$ $\times\mathbb{1}\left\{\frac{\mathrm{P}_k(V_j(s)|pa(V)_j(s))}{\mathrm{P}_j(V_j(s)|pa(V)_j(s))} \leq 2\log(2/\epsilon)M_{kj}\right\}$. Using Lemma 2 it is easy to observe that $\mathbb{E}[\tilde{Y}_k^\epsilon] \leq \mu_k \leq \mathbb{E}[\tilde{Y}_k^\epsilon] + \frac{\epsilon}{2}$ as $\eta_{kj}(\epsilon) \leq 2\log(2/\epsilon)M_{kj}$. Now, (24) together with this implies the lemma as $Z_k = \sum_{j\in[K]}\tau_j/M_{kj}$.

□

Theorem 4. *The estimator $\hat{Y}_k^\epsilon$ of (22) satisfies the following concentration guarantee:*

$$\mathbb{P}\left(\mu_k - \delta - \epsilon/2 \leq \hat{Y}_k^\epsilon \leq \mu_k + \delta\right) \geq 1 - 2\exp\left(-\frac{\delta^2\tau}{8(\log(2/\epsilon))^2}\left(\frac{Z_k}{\tau}\right)^2\right)$$

Proof. For the sake of analysis, let us consider the rescaled version $\bar{Y}_k^\epsilon = (Z_k/\tau)\hat{Y}_k^\epsilon$ which can be written as:

$$\bar{Y}_k^\epsilon = \frac{1}{\tau}\sum_{j=0}^{K}\sum_{s\in\mathcal{T}_j}\frac{1}{M_{kj}}Y_j(s)\frac{\mathrm{P}_k(V_j(s)|pa(V)_j(s))}{\mathrm{P}_j(V_j(s)|pa(V)_j(s))} \times\mathbb{1}\left\{\frac{\mathrm{P}_k(V_j(s)|pa(V)_j(s))}{\mathrm{P}_j(V_j(s)|pa(V)_j(s))} \leq 2\log(2/\epsilon)M_{kj}\right\}. \tag{25}$$

Since $Y_j(s) \leq 1$, we have every random variable in the sum in (25) bounded by $2\log(2/\epsilon)$

Let, $\bar{\mu}_k = \mathbb{E}[\bar{Y}_k^\epsilon]$. Therefore by Chernoff's bound, we have the following chain:

$$\mathbb{P}\left(|\bar{Y}_k - \bar{\mu}_k| \leq \delta\right) \leq 2\exp\left(-\frac{\delta^2\tau}{8(\log(2/\epsilon))^2}\right)$$

$$\implies \mathbb{P}\left(|\bar{Y}_k\frac{\tau}{Z_k} - \bar{\mu}_k\frac{\tau}{Z_k}| \leq \delta\frac{\tau}{Z_k}\right) \tag{26}$$

$$\leq 2\exp\left(-\frac{\delta^2\tau}{8(\log(2/\epsilon))^2}\right)$$

$$\implies \mathbb{P}\left(|\hat{Y}_k - \hat{\mu}_k| \leq \delta\frac{\tau}{Z_k}\right) \tag{27}$$

$$\leq 2\exp\left(-\frac{\delta^2\tau}{8(\log(2/\epsilon))^2}\right)$$

$$\implies \mathbb{P}\left(|\hat{Y}_k - \hat{\mu}_k| \leq \delta\right) \tag{28}$$

$$\leq 2\exp\left(-\frac{\delta^2\tau}{8(\log(2/\epsilon))^2}\left(\frac{Z_k}{\tau}\right)^2\right) \tag{29}$$

Now we can combine Equations (26) and (23) we get:

$$\mathbb{P}\left(\mu_k - \delta - \epsilon/2 \leq \hat{Y}_k^\epsilon \leq \mu_k + \delta\right) \geq 1 - 2\exp\left(-\frac{\delta^2\tau}{8(\log(2/\epsilon))^2}\left(\frac{Z_k}{\tau}\right)^2\right)$$

□

In Theorem 4, we observe that the first part of the exponent scales as $O(\epsilon^2\tau/(\log(2/\epsilon))^2)$ if we set $\delta = O(\epsilon)$, which is very close to the usual Chernoff's bound with τ i.i.d samples. The performance of this estimator therefore depends on the factor (Z_k/τ) which depends on the *fixed* quantities M_{kj} $(\forall j)$ and the allocation of the samples τ_j. In the next section, we will come up with a strategy to allocate the samples so that the estimators $\hat{Y}_k^\epsilon$ have good guarantees for all the arms k.

E.4. Allocation of Samples

In Section E.3, Theorem 4 tells us that the confidence guarantees on the estimator depends on how the samples are allocated between the arms. To be more precise, the term (Z_k/τ) in Equation (26), affects the performance of the estimator for μ_k $(\hat{Y}_k^\epsilon)$. We would like to maximize (Z_k/τ) for all arms $k \in [K]$.

Let the total budget be τ. Let $\mathcal{R}$ be the set of arms that remain in contention for the best optimal arm. Consider the matrix $\mathbf{A} \in \mathbb{R}^{K\times K}$ such that $\mathbf{A}_{kj} = 1/M_{kj}$ for all $k, j \in [K]$. Then, we decide the fraction of times arm k gets pulled, i.e. ν_k to maximize Z_k using the Algorithm 3.

Lemma 4. *Allocation $\boldsymbol{\tau}$ in Algorithm 3 ensures that $(Z_k/\tau) \geq \frac{1}{\sigma^*(B,\mathcal{R})}$ for all $k \in \mathcal{R}$.*

This is essentially the best allocation of the individual arm budgets in terms of ensuring good error bounds on the estimators $\hat{Y}_k^\epsilon$ for all $k \in \mathcal{R}$. Since **S1** is a special case of **S2**, to obtain the allocation for **S1** one needs to set the cost values c_i set to 1 for $i \in \mathcal{B}$ (*difficult* arms) in the above formulation and 0 otherwise.

E.5. Putting it together: Online Analysis

We analyze Algorithm 1 phase by phase. With some abuse of notation, we redefine various quantities to be used in the analysis of the algorithm. Each quantity depends on the phase indices, as follows:

- $\mathcal{R}(l)$: Set of arms remaining after phase $l-1$ ends.
- $\hat{Y}_k(l)$: The value of the estimator (in Algorithm 1) for arm k at the end of phase l.
- $\hat{Y}_H(l)$: The value of the highest estimate $\max_k \hat{Y}_k(l)$ (in Algorithm 1).
- $\mathcal{A}(l) \subseteq \mathcal{R}(l)$: Set of arms given by:

$$\mathcal{A}(l) := \left\{ k \in \mathcal{R}(l) : \Delta_k > \frac{10}{2^l} \right\}. \tag{30}$$

- S_l: Success event of phase l defined as:

$$S_l := \cap_{k \in \mathcal{R}(l), k \neq k^*} \left\{ \hat{Y}_k(l) \leq \mu_k + \frac{1}{2^{l-1}} \right\} \cap \left\{ \mu_k - \frac{3}{2^l} \leq \hat{Y}_{k^*}(l) \right\}. \tag{31}$$

Now we will establish that the occurrence of the event S_l implies that all arms in $\mathcal{A}(l)$ gets eliminated at the end of phase l, while at the same time the optimal arm survives. Consider an arm $k \in \mathcal{A}(l)$. Given S_l has happened we have:

$$\hat{Y}_H(l) \geq \hat{Y}_{k^*}(l) \geq \mu_{k^*} - \frac{3}{2^l}$$
$$\hat{Y}_k(l) \leq \mu_k + \frac{1}{2^{l-1}}$$

This further implies that $\hat{Y}_H(l) - \hat{Y}_k(l) \geq \Delta_k - 5/2^l > 5/2^l$. Therefore all the arms in $\mathcal{A}(l)$ are eliminated given S_l. Following similar logic it is also possible to show that the optimal arm survives. If $\hat{Y}_H(l) = Y_{k*}(l)$ then it survives certainly. Now, given S_l, only arms in $\mathcal{R}(l) \setminus \mathcal{A}(l)$ can be the ones with the highest means. Consider arms $k \in \mathcal{R}(l) \setminus \mathcal{A}(l)$. Again given S_l we have:

$$\hat{Y}_{k^*}(l) \geq \mu_{k^*} - \frac{3}{2^l}$$
$$\hat{Y}_k(l) \leq \mu_k + \frac{1}{2^{l-1}}$$

Therefore, we have $Y_k(l) - Y_{k^*}(l) \leq 5/2^l - \Delta_k < 5/2^l$. Therefore, the optimal arm survives.

It would seem that now it would be easy to analyze the probability of the event S_l, using Theorem 4. However, the bound in Theorem 4 depends on the sequence of arms eliminated so far in each phase. Therefore it is imperative to analyze $S_{1:l}$, that is the event that all phases from $1, 2, .., l$ succeed. Let $B_l = \mathbb{P}(S_l^c | S_{1:l-1})$. So, we have the chain:

$$\begin{aligned}
\mathbb{P}(S_{1:l}) &\geq \mathbb{P}(S_{1:l}|S_{1:l-1})\mathbb{P}(S_{1:l-1}) \\
&\geq \mathbb{P}(S_{1:l}|S_{1:l-1})\mathbb{P}(S_{1:l-1}|S_{1:l-2})\mathbb{P}(S_{1:l-2}) \\
&\geq \prod_{i=0}^{l-2} \mathbb{P}(S_{1:l-i}|S_{1:l-i-1})P(S_1) \\
&= \prod_{s=1}^{l} (1 - B_s) \\
&\geq 1 - \sum_{s=1}^{l} B_s
\end{aligned}$$

The advantage of analyzing the probability of $S_{1:l}$ is that given $S_{1:l}$ we know the exact sequences of the arms that have been eliminated till Phase l. This gives us exact control on the exponents in the bound of Theorem 4. Given $S_{1:s-1}$ we have,

$$\mathcal{R}(s) \subseteq \mathcal{R}^*(s) := \left\{ k : \Delta_k \leq \frac{10}{2^{s-1}} \right\}.$$

Recall that the budget for the samples of each arm in any phase s, is decide by solving the LP in Algorithm 3. Therefore, given $S_{1:s-1}$, we have $\sigma^*(B, \mathcal{R}(s)) \geq \sigma^*(B, \mathcal{R}^*(s))$. Therefore, we have the following key lemma.

Lemma 5. *We have:*

$$\begin{aligned}
B_l &:= \mathbb{P}\left(S_l^c | S_{1:l-1}\right) \\
&\leq 2|\mathcal{R}^*(l)| \exp\left(-\frac{2^{-2(l-1)} \tau(l) v^*(B, \mathcal{R}^*(l))^2}{8l^2} \right)
\end{aligned} \tag{32}$$

Proof. Note that in this phase we set $\eta_{kj} = 2lM_{kj}$. Setting $\epsilon = 2^{-(l-1)}$ and $\delta = 2^{-(l-1)}$ in Theorem 4 and by Lemma 4 we have:

$$\begin{aligned}
&\mathbb{P}\left(\mu_k - \frac{3}{2^l} \leq \hat{Y}_k(l) \leq \mu_k + \frac{1}{2^{l-1}} \right) \\
&\geq 1 - 2\exp\left(-\frac{2^{-2(l-1)} \tau(l) v^*(B, \mathcal{R}^*(l))^2}{8l^2} \right)
\end{aligned} \tag{33}$$

Note that the samples considered in phase l are independent of the event $S_{1:l-1}$. Doing a union bound of the event complementary to the success event in (31), for all the remaining arms in $\mathcal{R}^*(\ell)$ implies the result in the Lemma. □

Now we are at a position to introduce our main results as Theorem 5.

Theorem 5. *Consider a problem instance with K candidate arms $\{\mathrm{P}_k(V|pa(V)\}_{k=0}^{K-1}$. Let the gaps from the optimal arm be Δ_k for $k \in [K]$. Let us define the following important quantities:*

$$\mathcal{R}^*(\Delta_k) = \left\{ s : \left\lfloor \log_2\left(\frac{10}{\Delta_s}\right) \right\rfloor \geq \left\lfloor \log_2\left(\frac{10}{\Delta_k}\right) \right\rfloor \right\} \tag{34}$$

$$\bar{H}_k = \max_{\{l:\Delta_l \geq \Delta_k\}} \frac{\log_2(10/\Delta_l)^3}{(\Delta_l/10)^2 v^*(B, \mathcal{R}^*(\Delta_l))^2} \tag{35}$$

$$\bar{H} = \max_{k \neq k^*} \frac{\log_2(10/\Delta_k)^3}{(\Delta_k/10)^2 v^*(B, \mathcal{R}^*(\Delta_k))^2} \tag{36}$$

Algorithm 1 satisfies the following guarantees:

1. *The simple regret is bounded as:*

$$r(T,B) \leq 2K^2 \sum_{\substack{k\neq k^*: \\ \Delta_k \geq 10/\sqrt{T}}} \Delta_k \log_2\left(\frac{20}{\Delta_k}\right) \exp\left(-\frac{T}{2\bar{H}_k \overline{\log}(n(T))}\right) + \frac{10}{\sqrt{T}} \mathbb{1}\left\{\exists k \neq k^* \; s.t \; \Delta_k < 10/\sqrt{T}\right\}$$

2. *The error probability is bounded as:*

$$e(T,B) \leq 2K^2 \log_2(20/\Delta) \exp\left(-\frac{T}{2\bar{H}\overline{\log}(n(T))}\right)$$

The bound on the error probability only holds if $\Delta_k \geq 10/\sqrt{T}$ for all $k \neq k^*$.

Proof. Recall that the simple regret is given by:

$$r(T,B) = \sum_{k \neq k^*} \Delta_k \mathbb{P}\left(\hat{k}(T,B) = k\right) \tag{37}$$

Let us introduce some further notation. Let us define the phase at which an arm is *ideally* deleted as follows:

$$\gamma_k := \gamma(\Delta_k) := l \text{ if } \frac{10}{2^l} < \Delta_k \leq \frac{10}{2^{l-1}} \tag{38}$$

Therefore we have the following chain:

$$\begin{aligned}
&\mathbb{P}\left(\hat{k}(T,B) = k\right) \overset{a}{\leq} \mathbb{P}\left(S^{\mathrm{c}}_{1:\gamma_k}\right) \\
&\leq \sum_{l=1}^{\gamma_k} B_l \\
&\leq \sum_{l=1}^{\gamma_k} 2|\mathcal{R}^*(l)| \exp\left(-\frac{2^{-2(l-1)}\tau(l) v^*(B,\mathcal{R}^*(l))^2}{8l^2}\right)
\end{aligned}$$

provided $\Delta_k \geq 10/\sqrt{T}$. Justification for (a) - If arm k is chosen finally, it implies that it is not eliminated at phase γ_k. Therefore the regret of the algorithm is given by:

$$r(T,B) \leq \sum_{\{k\neq k^*:\Delta_k \geq 10/\sqrt{T}\}} \Delta_k \left(\sum_{l=1}^{\gamma_k} 2|\mathcal{R}^*(l)| \right. \tag{39}$$

$$\left. \exp\left(-\frac{2^{-2(l-1)}\tau(l) v^*(B,\mathcal{R}^*(l))^2}{8l^2}\right)\right) \tag{40}$$

$$+ \frac{10}{\sqrt{T}} \mathbb{1}\left\{\exists k \neq k^* \text{ s.t } \Delta_k < 10/\sqrt{T}\right\} \tag{41}$$

Let $\ell_1, \ell_2 .. \ell_s = \gamma_k$ such that $\mathcal{R}^*(\ell)$ changes value only at these phases. Let us set $\ell_{s+1} = \ell_s + 1$ for convenience in notation. Combining this notation with (39) we have:

$$r(T,B) \leq \sum_{\{k\neq k^*:\Delta_k \geq 10/\sqrt{T}\}} \Delta_k \left(\sum_{i=1}^{s} 2|\mathcal{R}^*(\ell_i)| \, (\ell_{i+1} - \ell_i) \right. \tag{42}$$

$$\left. \exp\left(-\frac{2^{-2\ell_i} T v^*(B,\mathcal{R}^*(l_i))^2}{2\ell_i^3 \overline{\log}(n(T))}\right)\right) \tag{43}$$

$$+ \frac{10}{\sqrt{T}} \mathbb{1}\left\{\exists k \neq k^* \text{ s.t } \Delta_k < 10/\sqrt{T}\right\}$$

Consider the phase ℓ_i when ideally at least an arm leaves. Let one of those arms be l. Recall that, γ_l is the phase where the arm ideally leaves according to (38). Therefore, $\gamma_l = \ell_i$. Also it is easy to observe that: $\mathcal{R}^*(\Delta_l) = \mathcal{R}^*(\ell_i)$. We have,

$$\ell_i \geq \log_2(10/\Delta_l) \tag{44}$$

as $\ell_{i+1} - \ell_i \leq \log_2(20/\Delta_k)$, $i \leq s$. Further, for every $\ell_i < \gamma_k$, there is at least one distinct arm $l : \gamma_l = \ell_i$. This is because an arm leaves only once ideally. Further, we associate ℓ_s with arm k although other arms may leave at the phase $\ell_s = \gamma_k$. Further, all arms l associated with $\ell_i < \gamma_k$ are such that $\Delta_l \geq \Delta_k$. This is because of (38) and the fact that $\ell_i < \ell_s = \gamma_k$. Therefore, the r.h.s in

Equation (42) is upper bounded as follows:

$$
\begin{aligned}
r(T,B) \leq & \sum_{\{k\neq k^*:\Delta_k\geq 10/\sqrt{T}\}} \Delta_k \left(\sum_{\{l:\Delta_l\geq\Delta_k\}} 2|\mathcal{R}^*(\Delta_l)| \right. \\
& \left. \log_2(20/\Delta_k)\exp\left(-\frac{(\Delta_l/10)^2 T v^*(B,\mathcal{R}^*(\Delta_l))^2}{2\log_2(10/\Delta_l)^3\overline{\log}(n(T))}\right)\right) \\
& + \frac{10}{\sqrt{T}}\mathbb{1}\left\{\exists k\neq k^* \text{ s.t } \Delta_k < 10/\sqrt{T}\right\} \\
\overset{(a)}{\leq} & \sum_{\{k\neq k^*:\Delta_k\geq 10/\sqrt{T}\}} \Delta_k \left(\sum_{\{l:\Delta_l\geq\Delta_k\}} 2|\mathcal{R}^*(\Delta_l)| \right) \\
& \log_2(20/\Delta_k)\exp\left(-\frac{T}{2\bar{H}_k\overline{\log}(n(T))}\right) \\
& + \frac{10}{\sqrt{T}}\mathbb{1}\left\{\exists k\neq k^* \text{ s.t } \Delta_k < 10/\sqrt{T}\right\} \\
\overset{(b)}{\leq} & 2K^2 \sum_{\{k\neq k^*:\Delta_k\geq 10/\sqrt{T}\}} \Delta_k \log_2(20/\Delta_k) \\
& \exp\left(-\frac{T}{2\bar{H}_k\overline{\log}(n(T))}\right) \\
& + \frac{10}{\sqrt{T}}\mathbb{1}\left\{\exists k\neq k^* \text{ s.t } \Delta_k < 10/\sqrt{T}\right\}
\end{aligned}
$$

Here (a) is by definition of $\bar{H}_k$ while (b) is because $|\mathcal{R}^*(\Delta_l)| \leq K$ and there are at most K terms in the summation.

Another quantity of interest here is the error probability. We will only provide bounds on the error probability $e(T,B)$ when we have $\Delta_k > 10/\sqrt{T}$ for all $k \neq k^*$. Let $\Delta = \min_{k\neq k^*}\Delta_k$ and $\gamma^* = \gamma(\Delta)$. In this case we have:

$$
\begin{aligned}
e(T,B) &\leq 1 - \mathbb{P}\left(S_{1:\gamma^*}\right) \\
&\leq \sum_{l=1}^{\gamma^*} B_l \qquad (45) \\
&\leq \sum_{l=1}^{\gamma^*} 2|\mathcal{R}^*(l)| \exp\left(-\frac{2^{-2(l-1)}\tau(l) v^*(B,\mathcal{R}^*(l))^2}{8l^2}\right) \\
&\overset{a}{\leq} \sum_{l=1}^{\gamma^*} 2|\mathcal{R}^*(l)| \exp\left(-\frac{2^{-2l} T v^*(B,\mathcal{R}^*(l))^2}{2l^3\overline{\log}(n(T))}\right) \qquad (46)
\end{aligned}
$$

(a)- This follows from the definition of $\tau(l)$.

As before let $\ell_1 = 1, \ell_2..\ell_m \leq \gamma^*$ such that $R^*(\ell)$ changes value only at these phases. Let us set $\ell_{m+1} = \ell_m + 1$ for convenience in notation. Then, $e(T,B)$ in (46) is upper bounded by:

$$
\begin{aligned}
e(T,B) \leq & \sum_{i=1}^{m} 2|\mathcal{R}^*(\ell_i)|\left(\ell_{i+1}-\ell_i\right) \qquad (47) \\
& \exp\left(-\frac{2^{-2\ell_i} T v^*(B,\mathcal{R}^*(l_i))^2}{2\ell_i^3\overline{\log}(n(T))}\right)
\end{aligned}
$$

Consider the phase ℓ_i when ideally at least an arm leaves. Let one of those arms be k. Recall that, γ_k is the phase where the arm ideally leaves according to (38). Therefore, $\gamma_k = \ell_i$. Also it is easy to observe that: $\mathcal{R}^*(\Delta_k) = \mathcal{R}^*(\ell_i)$. Then,

$$\ell_i \geq \log_2(10/\Delta_k) \qquad (48)$$

Further, for every ℓ_i, there is a distinct and different k : $\gamma_k = \ell_i$. This is because an arm leaves only once ideally. Therefore, according to (48) and (47) we have:

$$
\begin{aligned}
e(T,B) \leq & \sum_{k\neq k^*} 2|\mathcal{R}^*(\Delta_k)|\gamma^* \qquad (49) \\
& \exp\left(-\frac{(\Delta_k/10)^2 T v^*(B,\mathcal{R}^*(\Delta_k))^2}{2\log_2(10/\Delta_k)^3\overline{\log}(n(T))}\right) \\
& \leq 2K^2\log_2(20/\Delta)\exp\left(-\frac{T}{2\bar{H}\overline{\log}(n(T))}\right) \qquad (50)
\end{aligned}
$$

Here, we have used the definition of $\bar{H}$ and the fact that $|\mathcal{R}^*(\Delta_k)| \leq K$ and $\gamma^* \leq \log(20/\Delta)$ □

F. More Experiments

In this section we provide more details about our experiments. We also include extensive synthetic simulations results.

F.1. Synthetic Experiments

In this section, we empirically validate the performance our algorithm through synthetic experiments. We carefully design our simulation setting which is simple, but at the same time sufficient to capture the various tradeoffs involved in the problem. An important point to note is that our algorithm is not aware of the actual effect of the changes on the target (gaps between expectations) but it only knows the divergence among the candidate soft interventions. Sometimes, a change with large divergence from an existing one may not maximize the effect we are looking for. Conversely, smaller divergence may sometimes lead you closer to the optimal. We demonstrate that our algorithm performs well in all the experiments, as compared to previous works (Audibert & Bubeck, 2010; Lattimore et al., 2016).

Experimental Setup: We set up our experiments according to the simple causal graph in Figure 6. V is assumed to

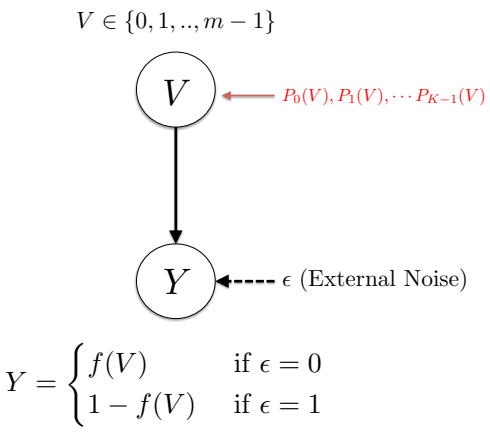


Figure 6. Causal Graph for Experimental Setup

be a random variable taking values in $\{0, 1, 2, \cdots, m-1\}$. The various arms $\mathrm{P}_0(V), \mathrm{P}_1(V), ...\mathrm{P}_{K-1}(V)$ are discrete distributions with support $[m]$. We will vary m and K over the course of our experiments.

Y is assumed to be a function of V and some random noise ϵ which is external to the system. In our experiments, we set the function as follows:

$$Y = \begin{cases} f(V) & \text{if } \epsilon = 0 \\ 1 - f(V) & \text{if } \epsilon = 1 \end{cases}$$

where $f : [m] \to \{0, 1\}$ is an arbitrary function. We set $\mathbb{P}(\epsilon = 1) = 0.01$ in all our experiments. The discrete candidate distributions are modified to explore various trade-offs between the gaps and the effective standard deviation parameters.

Budget Restriction: The experiments are performed in the budget setting **S1**, where all arms *except* arm 0 are deemed to be *difficult*. We plot our results as a function of the total samples T, while the fractional budget of the *difficult* arms (B) is set to $1/\sqrt{T}$. Therefore, we have $\sum_{k \neq 0} T_k \leq \sqrt{T}$. This essentially belongs to the case when there is a lot of data that can be acquired for a default arm while any new change requires significant cost in acquiring samples.

Competing Algorithms: We test our algorithms on different problem parameters and compare with related prior work (Audibert & Bubeck, 2010; Lattimore et al., 2016). We briefly describe the algorithms compared:

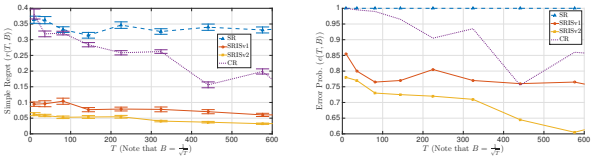
1. **SRISv1:** This is Algorithm 1 introduced in Section 3.2.
2. **SRISv2:** This is Algorithm 2 which is a simple modification of SRISv1, as detailed in Section 3.2.
3. **SR:** This is the best arm identification algorithm from (Audibert & Bubeck, 2010) adapted to the budget setting. The division of the total budget T into $K - 1$ phases is identical, while the individual arm budgets are decided in each phase according to the budget restrictions.
4. **CR:** This is Algorithm 2 from (Lattimore et al., 2016). The optimization problem for calculating the mixture parameter $\boldsymbol{\eta}$ has been modified to account for the budget restrictions. This is a natural modification to the algorithm.

Experiments: In our experiments, we choose f to be the parity function, when $V \in [m]$, is represented in base 2. Note that arm 0 is the arm that can be sampled $O(T)$ times while the rest of the arms can only be sampled $O(\sqrt{T})$ times due to the above budget constraints. So, the divergence of the arm 0 from other arms is crucial alongside the gaps. We perform our experiments in different regimes that get progressively easier. In these experiments, we function in various regimes of the divergences between the other arms and arm 0, and the gaps from the optimal arm in terms of target value. When there is no information leakage, the samples are divided among the K arms. So, the loss in having multiple arms can be expressed as a scaling $\sqrt{K}$ in standard deviation. Recall the log divergence measure M_{k0} which is a measure of information leakage from arm 0 to another arm k. Therefore, in the following, when we say high divergence from arm 0, it means that $M_{k0}/\sqrt{K}$ is high for most arms $k \neq 0$.

High Divergence and Low Δ: This is the hardest of all settings. Here, we set $m = 20$ and $K = 30$. Here, we have M_{k0} to be pretty high for all the arms $k \neq 0$. This means that the arm 0, which can be pulled $O(T)$ times provides highly noisy estimates for other arms. We have $M_{k0}/\sqrt{K} \sim 30$ for most arms. Moreover, the minimum gap from the best arm $\Delta = 0.04$, which is pretty small. This implies that it is harder to distinguish the best arm.

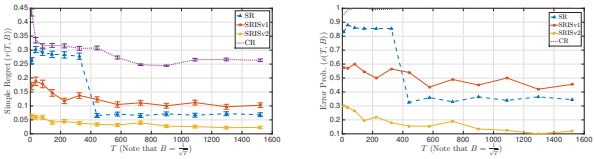
The results are demonstrated in Figure 7. Figure 7a displays the simple regret. We see that both SRISv1 and SRISv2 outperform the others by a large extent, in this hard setting, even when the number of samples are very low. In Figure 7b we plot the probability of error in exactly identifying the best arm. We see that none of the algorithms successfully identify the best arm, in the small sample regime, as the gap Δ is very low. However, our algorithms quickly zero in on arms that are *almost* as good as the optimal, and therefore the simple regret is well-behaved. Our algorithm performs this well even when the divergences are big, because it is able to reject the arms that have high Δ_i in the early phases, very effectively.

High Divergence and High Δ: This is easier than the pre-



(a) Simple Regret (b) Probability of Error

Figure 7. Performance of various algorithms when divergences M_{k0}'s are high and mininum gap Δ is small. The results are averaged over the course of 500 independent experiments. The total sample budget T is plotted on the x-axis. Note that budget for all arms other than arm 0 is constrained to be less than $\sqrt{T}$. Here $K = 30$.



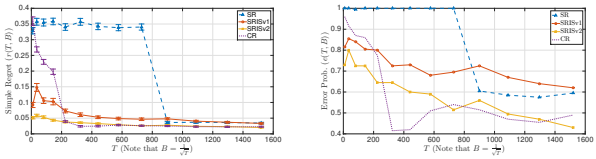
(a) Simple Regret (b) Probability of Error

Figure 8. Performance of various algorithms when divergences M_{k0}'s are high and min. gap Δ is not too bad. The results are averaged over the course of 500 independent experiments. The total sample budget T is plotted on the x-axis. Note that budget for all arms other than arm 0 is constrained to be less than $\sqrt{T}$. Here, $K = 20$.

vious setting. Here, we set $m = 10$ and $K = 20$. Here, we have M_{k0} to be very high for all the arms $k \neq 0$. Thus arm 0 provides very noisy estimates on other arms. We have $M_{k0}/\sqrt{K} \gg 50$ for many arms. However, the minimum gap from the best arm $\Delta = 0.15$, which is not too small. This implies that it might be easier to distinguish the best arm.

The results are demonstrated in Figure 8. Figure 8a displays the simple regret. We see that in the small sample regime SRISv1 and SRISv2 outperform the others by a large extent. In the high sample regime, SRISv2 is still the best, while SR and SRISv1 are close behind. In Figure 8b we plot the probability of error in exactly identifying the best arm. We see that SRISv2 performance very well in identifying the best arm even though arm 0 gives highly noise estimates. It is interesting to note that CR does not perform well. This can be attributed to the non-adaptive clipper in CR, that incurs a significant bias because arm 0 has high-divergences from most of the other arms.

Low Divergence and Low Δ: This is another moderately hard setting, similar to the previous one. Here, we set $m = 20$ and $K = 30$. Here, we have M_{k0} to be not too high for the arms $k \neq 0$. This means that the arm 0, which can be pulled $O(T)$ times is moderately good for estimating the other arms. Here, $M_{k0}/\sqrt{K} \leq 10$ for most arms k. However, the minimum gap from the best arm $\Delta = 0.04$, which is small. This implies that it might be hard to distinguish the best arm.



(a) Simple Regret (b) Probability of Error

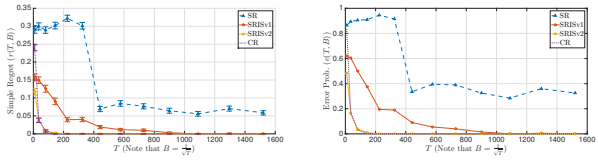
Figure 9. Performance of various algorithms when divergences M_{k0}'s are moderately low and min. gap Δ is small. The results are averaged over the course of 500 independent experiments. The total sample budget T is plotted on the x-axis. Note that budget for all arms other than arm 0 is constrained to be less than $\sqrt{T}$. Here, $K = 30$.

The results are demonstrated in Figure 9. Figure 9a displays the simple regret. We see that in the small sample regime SRISv1 and SRISv2 outperforms the others by a large extent. In the high sample regime, SRISv2 is still the best, while CR is close behind. In Figure 9a we plot the probability of error in exactly identifying the best arm. We see that most of the algorithms have moderately bad probability of error as the gap Δ is small. However, the algorithms SRISv2 and SRISv1 are quickly able to zero down on arms close to optimal as shown in the simple regret in the small sample regime.

Low Divergence and High Δ: This is the easiest of all settings. Here, we set $m = 10$ and $K = 20$. Here, arms 0 has $\mathrm{P}_0(V)$ pretty close to the uniform distribution on $[m]$. Therefore, it is very well-posed for estimating the means of all other arms. In fact we have $M_{k0}/\sqrt{K} < 2$ for many arms. Moreover, the minimum gap from the best arm $\Delta = 0.15$, which is not too small. This implies that it might be very easy to distinguish the best arm.

The results are demonstrated in Figure 10. Figure 10a displays the simple regret. We see that SRISv2 and CR perform extremely well closely followed by SRISv1. In Figure 10b we plot the probability of error in exactly identifying the best arm. Again SRISv2 and CR have almost zero probability of error and SRISv1 is close behind. This is because Δ is pretty large. In this example, we observe that all the algorithms that use information leakage are better than SR, because arm 0 is well-behaved. CR performs almost as well as SRISv2 in this example, as the static clipper is never invoked because almost always the ratios in the importance sampler are well bounded.

In conclusion, it should be noted that our algorithms perform well in all the different settings, because they are able



(a) Simple Regret (b) Probability of Error

Figure 10. Performance of various algorithms when divergences M_{k0}'s are low and min. gap Δ is not too bad. The results are averaged over the course of 500 independent experiments. The total sample budget T is plotted on the x-axis. Note that budget for all arms other than arm 0 is constrained to be less than $\sqrt{T}$. Here $K = 20$.

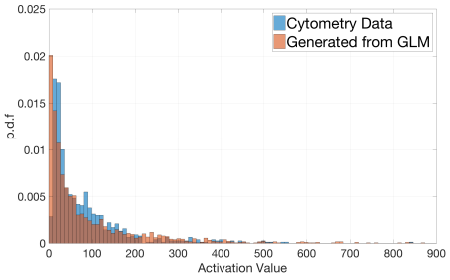


Figure 11. Histograms of data from the cytometry data-set and from the GLM trained for the activations of an internal node *pip2*.

to adapt to the problem parameters (similar to (Audibert & Bubeck, 2010)) and at the same time leverage the information leakage (similar to (Lattimore et al., 2016)).

F.2. More on Flow Cytometry Experiments

In this section we give further details on the flow cytometry experiments. As detailed in the main paper we use the causal graph in Fig. 5(c) in (Mooij & Heskes, 2013) (shown in Fig. 3a) as the ground truth. Then we fit a GLM gamma model (Hardin et al., 2007) between each node in the graph and its parents using the observational data-set. The GLM model produces a highly accurate representation of the flow cytometry data-set. In Fig. 11 we plot the histogram for the activation of an internal node *pip2* from the real data and samples generated from the GLM probabilistic model. It can be seen that the histograms are very close to each other.

In Fig. 12 we plot the performance of SRISv2 when the divergence metric is replaces by KL-divergence. In one of the plots SRISv2 is modified by setting $M_{ij} = 1 + KL(P_i, P_j)$. It can be seen that the performance degrades, which signifies that our divergence metric is fundamental to the problem.

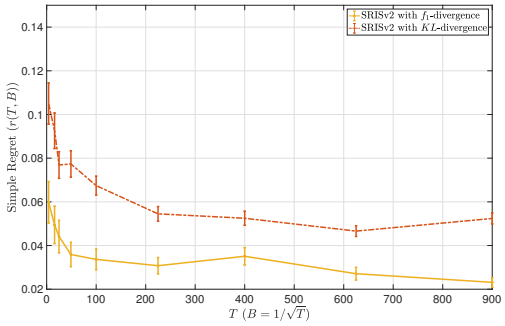


Figure 12. Comparison of SRISv2 with two different divergence metric. It shows that our divergence measure is fundamental for good performance. The experiments have been performed in a setting identical to Fig. 3c

F.3. More on Interpretation of Inception Deep Network

In this section we describe the methodology of our model interpretation technique in more detail. In Section 4.2 we have described how the best arm algorithm can be used to pick a distribution over the superpixels of an image, that has the maximum likelihood of producing a certain classification from Inception. Here, we describe how the distributions over the superpixels are generated and how they are used subsequently. The arm distributions are essentially points in the n-dimensional simplex (where n is the number of superpixels into which the image is segmented). These distributions are generated in a randomized fashion using the following methods:

1. Generate a point uniformly at random in the n-dimensional simplex.
2. Randomly choose $l < n$ superpixels. Make the distribution uniform over them and 0 elsewhere.
3. Randomly choose $l < n$ superpixels. The probability distribution is a uniformly chosen random point over the l-dimensional simplex with support on the l chosen superpixels and 0 elsewhere.
4. Start a random walk from a few superpixels which traverses to adjacent superpixels at each step. Stop the random-walk after a certain number of steps and choose the superpixels touched by the random walk. Then choose a uniform distribution over the super pixel support or choose a random distribution from the simplex of probability distributions over the support of the chosen superpixels (like in the previous point) and 0 elsewhere. This method uses the geometry of the image.

Note that all the above methods do not depend on the specific content of the images. Using the above methods L

pull arms are chosen which are used to collect the rewards. Further there are K *opt* arms that are the interventions to be optimized over. When an arm is used to sample, then $m \ll n$ superpixels are chosen with replacement from the distribution of that arm. These pixels are preserved in the original image while everything else is blurred out before feeding this into the neural network. Thus if the distribution corresponding to an arm is P, then the actual distribution to be used for the importance sampling is the product distribution P^m. Note that the *pull* arms are separate from the *opt* arms. The true counterfactual power of our algorithm is showcased in this experiment, as we are able to optimize over a large number of interventions that are never physically performed.