Learning to Exploit Exploration

Allan Axelrod, *Student Member, IEEE,* and Dr. Girish Chowdhary, *Member, IEEE*

***Abstract*—The explore-exploit dilemma, i.e., when it is best to explore (learn) or exploit (optimize), is a problem of critical interest for Bandit, Reinforcement Learning, and Learning-in-Control formulations. For such formulations, an agent must learn how to optimally solve a problem, and learning may be a persistent requirement if the environment dynamics are changing. As learning actions are inherently less cost effective than optimization actions, a learning policy is considered optimal if it obtains the highest information gain per learning action, so ideally less learning actions would be required before optimization actions can be taken. However, we show that an optimal learning policy is not generally attainable without sufficient sampled information. Consequently, we theoretically derive criteria to identify when our agent can reliably select optimal learning actions. A so-called uninformed-to-informed algorithm, initialized with some prior learning policy, allows our agent to select optimal learning actions once our criteria is satisfied. We validate our algorithm on 4 real-world spatiotemporal datasets against baseline algorithms and show that our algorithm learns to exploit exploration.**

***Index Terms*—Delay Systems, Learning, Maximum Entropy Methods, Sampling Methods, Time-varying Channels.**

I. Introduction

The certainty equivalence property is extensively used in literature pertaining to linear analysis and control in the presence of additive noise **certainty equivalence citations here**, as solutions to their deterministic counterpart are also optimal for the stochastic case. Yet the vast majority of practical systems, such as systems with uncertain parameters or nonlinear systems, are not amenable to the certainty equivalence property **non CEQ citations here**. For this broad class of stochastic systems, an optimal solution to the deterministic counterpart of a control problem is not optimal for the stochastic system. Consequently, the accuracy with which we can estimate stochastic uncertainties significantly affects the performance of stochastic control systems.

Sample-based probability bounds are widely used for ensuring accurate stochastic system estimation **sample-based bound citations here**. Once a sufficient number of samples have been obtained, so as to satisfy a sample-based probability bound, it is assumed that the sample-based probability bound is then satisfied for all time. This assumption is valid if the stochastic process is modeled correctly. That is to say, if a random variable is assumed to be independent and identically distributed (i.i.d.) or if a time-varying stochastic process has parameters that evolve as prescribed by the designer for all time, then a sample-based probability bound, once satisfied, should hold for all future samples.

In this note, we show that if a stochastic process is partially observable, then the accuracy of a sample-based model of the stochastic process may depreciate with respect to the time between consecutive samples. This is relevant to stochastic systems whereby the form of the stochastic system is not known *a priori* and is critically important for systems with delay times which may depend on the control system inputs.

We develop a theoretical framework for learning the rates at which the accuracy of each sample-based model in a sensor network depreciates over time in terms of the information gain [2], we call this rate the *information exposure rate*. Our aim is to leverage the learned information exposure rates to develop an informed exploration (i.e., information-driven channel selection) policy which prioritizes our limited sensing authority based on the expected information gain of each channel to optimize the accuracy of our sensor network.

We perform validation experiments using 4 real-world datasets including the Intel Berkeley indoor temperature dataset [11], the European Research Area daily interim outdoor temperature dataset [12], the Washington rainfall dataset [15], and the Ireland windspeed dataset [14]. Our initial results show that sequential and random exploration policies outperform both our proposed and an existing informed exploration policy called predicted information gain (PIG) [3] in terms of their ability to select the subset of channels which result in the highest information gain, this is evaluated offline.

Our initial results motivate our derivation and implementation of sample-based bounds on the information exposure rate in a so-called *uninformed-to-informed* exploration policy; i.e., a policy which initializes as an uninformed policy, such as sequential or random exploration, and then transitions to an informed exploration policy once a sample-based bound on the information exposure rate is satisfied. Our final results show that our proposed uninformed-to-informed exploration policy outperforms sequential, random and PIG exploration policies in terms of their ability to select the subset of channels which result in the highest information gain.

II. Formulation

We desire to learn a policy such that learning from data is optimized, as shown in Figure **??**. Specifically, we wish to learn about a process $\mathcal{Y}$, which could be stationary or nonstationary, that is dependent on an unobserved change-inducing process $\mathcal{X}$. We learn a predictive model on the mutual dependence between $\mathcal{Y}$ and $\mathcal{X}$ using their mutual information. Our predictive model on the mutual information between $\mathcal{X}$ and $\mathcal{Y}$ allows us to *anticipate* where samples of $\mathcal{Y}$ will be the most affected by the change-inducing process $\mathcal{X}$. As Section II-A will show, the resultant model is equivalent to modeling the entropy injected into $\mathcal{Y}$ by $\mathcal{X}$.

A. Axelrod and Dr. G. Chowdhary were with the Department of Mechanical and Aerospace Engineering, Oklahoma State University, Stillwater, OK, 74074 USA e-mail: 213axelrod@gmail.com & girish.chowdhary@okstate.edu.

Manuscript received January 19, 2016; revised TBD 2016.

A. Preliminaries on Mutual Information

Mutual information, which quantifies the mutual dependence between process $\mathcal{X}$ and the process we seek to learn, $\mathcal{Y}$, is defined as

$$I(X;Y) = D_{KL}(p(x,y)||p(x)p(y)), \tag{1}$$

where $D_{KL}(p(x,y)||p(x)p(y))$ is the Kullback-Leibler (KL) divergence, which quantifies the information contained in $p(x,y)$ that is not in $p(x)p(y)$ as

$$D_{KL}(\hat{q}||\hat{p}) = \int_{-\infty}^{\infty} \hat{q} \ln \frac{\hat{q}}{\hat{p}} dx, \tag{2}$$

where $\hat{q}$ and $\hat{p}$ are the beliefs $p(x,y)$ and $p(x)p(y)$, respectively.

As a result of (1), we know that mutual information is symmetric; i.e., $I(X;Y) = I(Y;X)$. Mutual information can be equivalently represented as the value for a function of a random variable. This makes it feasible to learn a predictive model on Kullback-Leibler (KL) divergence for informed exploration as

$$\begin{aligned} I(Y;X) &= \textstyle\int_X p(x) D_{KL}(p(y|x)||p(y))dx \\ &= \mathbb{E}_X(D_{KL}(p(y|x)||p(y))). \end{aligned} \tag{3}$$

Index of Notation	
K	Number of Bandit Arms
η_t	Selected Subset of Arms at Time t
η	History of Selected Arms
i	Index of Arm
$\tau_{(i,n)}$	Timestamp of n^{th} sample at Arm i
t	Time $t \in [\tau_i^n, \tau_i^{n+1}]$
κ	Plays per Episode where $Car(\eta) \leq K$
Z_i	D_{KL}(Posterior||Prior) at Arm i
$\Lambda_i(\cdot)$	Exposure at Arm i
λ_i	Homogeneous Exposure Rate at Arm i
$m_i(\cdot)$	Gaussian Process (GP) Mean at Arm i
$k_i(\cdot,\cdot)$	Covariance Kernel of GP at Arm i
α_i	Gamma Distribution Shape at Arm i
β_i	Gamma Distribution Rate at Arm i
Δt_i	Time Between Samples at Arm i
σ_i^2	Predictive Confidence of GP at Arm i
$\sigma^2_{(i,\hat{q})}$	Variance of Prior at Arm i
$\sigma^2_{(i,\hat{q})}$	Variance of Posterior at Arm i
$\lambda_{r,i}(\cdot)$	Randomly Seeded Pep Parameter
ϵ	Gaussian White Noise

Algorithm 1: Uninformed-to-Informed RAPTOR

```
Input:$\kappa,\ K,\ c,\ m,\ k$
Output:$\eta$
$(\alpha_{(i,0)}, \beta_{(i,0)}, n_{(i,0)}, \eta_{\tau_0}, \tau_{(i,0)}, \Delta t_i) \leftarrow 0 \ \forall \ i$
for each episode at time $t$ do
    for each $i \in S$ do
        if $n_i \geq 1$ then
            $\Delta t_i \leftarrow t - \tau_{(i,n_i)}$
        if Update of (9)$< 0$ then
            $\mathbb{E}(Z_i(t)) = Z_i(\tau_{(i,n)}) + \lambda_i \Delta t_i$
    $b \leftarrow \underset{i}{\text{argmax}} \ \frac{1}{n_i \cdot (\tau_{(i,n)} - \tau_{(i,1)})\sqrt{\lambda_i}}$
    if $\tau_{(b,n)} > \tau_{(b,1)} + \frac{1}{n_b\sqrt{\lambda_b}}$ (Corollary IV.6) then
        $\eta_t \leftarrow \underset{\eta \subset \{1,\ldots,K\}}{\text{argmax}} \ \mathbb{E}\Big[\sum_{i\in\eta} Z_i\Big]$
        s.t. $\text{Car}(\eta) \leq \kappa$
    else
        Uninformed Exploration (Sequential, Random,
        etc.)
    for each $i \in \eta_t$ do
        $(\mu_{(i,\hat{q})}, \sigma^2_{(i,\hat{q})}) \leftarrow$Update (21)
        $Z_i(t) \leftarrow$Update (22)
        if $\Delta Z_i \leftarrow Z_i(t) - Z_i(\tau_{(i,n_i)}) < 0$ then
            $\Delta Z_i \leftarrow 0$
        $\tau_{(i,n_i+1)} \leftarrow t$
        $n_i \leftarrow 1 + n_i$
        $\alpha_i \leftarrow \alpha_i + \Delta Z_i$
        $\beta_i \leftarrow \beta_i + \Delta t_i$
        $\lambda_i \leftarrow \frac{\alpha_{(i,0)} + \alpha_i}{n_{(i,0)}\beta_{(i,0)} + n_i\beta_i}$
        $GP_i(m_i(\Delta t_i), k_i(\Delta t_i, \Delta t_i')) \leftarrow [\Delta t_i, \Delta Z_i]$
```

B. Modeling Mutual Information

As measurements of the change-inducing process, $\mathcal{X}$, are not available, we use the following hierarchical structure to model the effect of $\mathcal{X}$ on $\mathcal{Y}$

$$\begin{aligned} x &\sim \mathcal{X} \\ y &\sim \mathcal{Y}|x \\ z &\sim \mathcal{Z}|x,y \ , \end{aligned} \tag{4}$$

where $z = D_{KL}(p(y|x)||p(y))$ is a mapping of the entropy injected into posterior of $\mathcal{Y}$ by $\mathcal{X}$; i.e., z is a mapping of interesting change in the process we are learning, $\mathcal{Y}$, due to $\mathcal{X}$. We model $\mathcal{Z}$ as a Poisson exposure process (Pep), formally introduced in Section II-C, with a parameter Λ. We learn an expectation on the Pep in (4) to *anticipate* regions where $\mathcal{Y}$ is most heavily affected by the unobserved process $\mathcal{X}$; i.e.,

$$I(Y;X) \approx \mathbb{E}_Z(D_{KL}(p(y|x)||p(y))). \tag{5}$$

Admittedly, if $p(x) \neq p(z)$ in (5), the approximation would be of questionable quality. However, we mitigate this concern by developing a probabilistic error bound on our expectation of $D_{KL}(p(y|x)|p(y))$ in Section IV, based on the exposure of our agent to the sensing domain.

C. Poisson Exposure Process Model

Definition II.1.
The Poisson exposure process (Pep) is defined as

$$f(z|\Lambda(t)) = C_{\Lambda(t)} \frac{(\Lambda(t))^z e^{-\Lambda(t)}}{\Gamma(z+1)}, \tag{6}$$

where $\Lambda(t) = \lambda t$ is a homogeneous Pep with an exposure rate of λ, C_λ is the normalizing constant, and $Z(\tau^i)$ is termed the i^{th} Poisson exposure trial. When $\Lambda(t) \neq \lambda t \ \forall \ t$, the Pep is termed inhomogeneous. We term the Pep as a Poisson exposure distribution (Ped) when $\Lambda(t)$ is a constant, as in [4], [5].

In [4], [5], the maximum likelihood estimate of the Poisson exposure distribution, and therefore the Pep by extension, was found to be highly nonlinear. However, as Definition II.1 shows that the Pep is similar in form to the Poisson process in (6), we show that the conjugate prior of the Pep is identical to that of the Poisson process as shown in Fact II.2.

Fact II.2.
The gamma distribution is a conjugate prior of the homogeneous Poisson exposure process (Pep) such that

$$G\left(\lambda^{*}t|\alpha+z,\beta+t,z\right)\propto Pep\left(z|\lambda t\right)G\left(\lambda t|\alpha,\beta\right). \tag{7}$$

Proof. See Appendix B for details. □

Fact II.2 provides us a simple analytical update for the homogeneous Pep.

1) Cox Gaussian Process: A general model for spatiotemporally varying Poisson processes is the Cox Process. The intensity parameter, λ, of a Cox process is called doubly-stochastic as the model parameter, Λ, is itself drawn from another stochastic process. However, existing Cox Process models require a priori output scaling or domain specification [6], [7], [8], [9]. Since an upper bound on Λ may not be known a priori, we introduce a novel approximation of the Bayesian Nonparametric model termed the *Cox Gaussian Process (CGP)*, which models the KL divergence growth as $Z \approx= \int_{\tau_n}^{t}\Lambda(t)dt$ using the following Gaussian Process prior

$$\begin{aligned}\lambda\Delta t &\sim Pep\left(\int_{\tau_n}^{t} g\left(\Lambda(t)\right)dt\right)\\ \int_{\tau_n}^{t} g\left(\Lambda(t)\right)dt &\sim GP(m(t),k(t,t')),\end{aligned} \tag{8}$$

where $Pep(\cdot)$ is the Poisson exposure process, $\Lambda(t)$ is the exposure at time t, λ is the homogeneous exposure rate at time t, and GP is a Gaussian Process with the mean m and covariance kernel $k(\cdot,\cdot)$. Although a Gaussian Process is technically too general because it can model negative values as well, we train the Gaussian process on $D_{KL}(\cdot)\geq 0$ and we bound the output of the Gaussian process so that $\int_{\tau_n}^{t}\Lambda(t)dt \geq 0$.

2) Poisson-Cox Gaussian Process: As will be shown in Section IV, an exposure bound on the estimation error of the homogeneous (linear) Pep may be used to switch to informed exploration from uninformed exploration. In order to capitalize on this bound, our approach must consider the homogeneous Pep in some facet. Below, we present the Poisson-Cox Gaussian process, which initializes as a Pep and then smoothly transitions to a Cox Gaussian process.

$$\begin{aligned}\mathbb{E}[Z|\Delta t] \leftarrow{}& (1-\sigma^2)\mathbb{E}_{GP}(\Delta Z|\Delta t)\\ &+(\sigma^2)\mathbb{E}_{Pep}(\Delta Z|\Delta t)+Z(\tau_n),\end{aligned} \tag{9}$$

where $\sigma^2 \in [0,1]$ is the predictive confidence of the Gaussian process, where σ^2 is initially 0 and σ^2 varies as a function of the samples obtained.

As we show in Section III, a bound is still needed to mitigate the extended duration of inefficient exploration that results from engaging in informed exploration using (9) with from the onset of informed exploration experiments. Using (9), the bound for the homogeneous Pep may be leveraged, and a smooth transition from the linear regression to the doubly stochastic and nonlinear regression is also facilitated.

D. Multi-Play N-Armed Bandit Formulation

As noted in Section **??**, we consider our sensing domain as a multi-play n-armed bandit problem. We let K denote the number of bandit arms in our sensing domain. We assume that the KL divergence at each arm i may be modeled as an independent process $\mathcal{Z}_i$. We let the random variable $Z_i(\tau_i^n)$ denote the KL divergence obtained at arm i, where $\tau_{(i,n)}$ is the time of the n^{th} visit to arm i. At each time instance (i.e., episode), our agent is capable of pulling a subset of bandit arms, which can be thought of as equivalent to visiting a sensing location or sensor. We denote these arms by $\eta \subset \{1,2,\ldots,K\} \ni Car(\eta) \leq \kappa$, where $\mathrm{Car}(\eta)$ denotes the cardinality of the index set η.. The objective of the optimal informed exploration policy is to maximize the total KL divergence obtained in each episode. Hence, we aim to solve:

$$\begin{aligned}\eta^{*} := \arg\max_{\eta\subset\{1,\ldots,K\}} \quad & \mathbb{E}\Big[\sum_{i\in\eta} Z_i\Big]\\ \text{subject to} \quad & \mathrm{Car}(\eta)=\kappa.\end{aligned}$$

III. Informed Exploration Results

In directly applying the policy in Section II-D, where $\mathbb{E}(Z_i)$ is calculated using (9), we see extended sampling durations of poor performance for informed exploration across a range of allowed plays per episode (κ). For these experiments, we model $\mathcal{Y}$ using (21) and $z \sim \mathcal{Z}$ is calculated using (22).

Figure 2 shows a clear relationship to the optimality of informed policies and the exposure of the agent to the dataset. In Section IV, we formalize this relationship as a bound for uninformed-to-informed exploration.

IV. Exposure Bound

Exposure bounds are defined in terms of the number of samples and the time between the first and most-recent samples. Exposure bounds intuit that a number of samples over long durations contain more information than the same number of samples over short durations in varying domains. Hence, lower exposure rates lead to a greater sampling time requirement for the transition between uninformed and informed exploration.

A. Preliminary on Probability Inequalities

We start with the Chebyshev inequality, shown in Lemma IV.1, and we then consider the Bienaymé-Chebyshev inequality, shown in Lemma IV.2. In using our model on the Pep in Lemma IV.1, we show that our exposure bound provides a similar utility to that of the Bienaymé-Chebysehv inequality, but where the independent and identically distributed (iid) requirement in the Bienaymé-Chebyshev inquality is reduced to an independence requirement for our exposure bound.

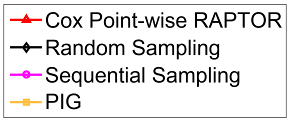
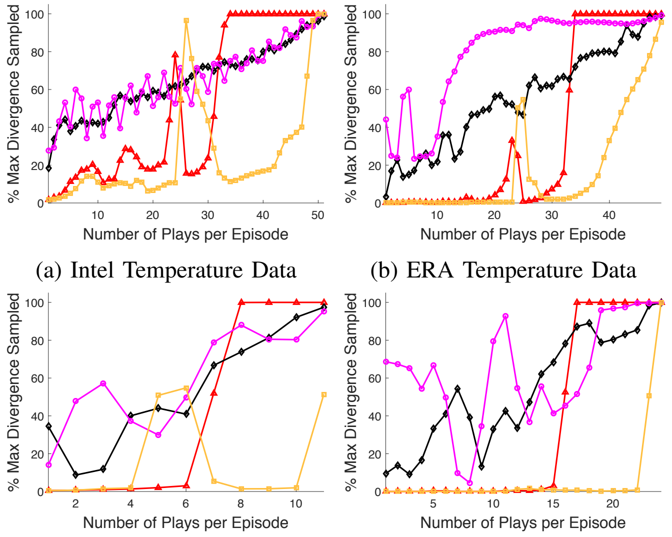


Figure 1: A legend of RAPTOR and baseline methods.



(a) Intel Temperature Data (b) ERA Temperature Data

(c) Ireland Windspeed Data (d) Washington Rain Data

Figure 2: The informed exploration parameters for PIG and RAPTOR are initialized using 2 sequential sampling sweeps across bandits. Experiments are performed on stationary and nonstationary datasets.

Lemma IV.1 (Chebyshev Inequality). *Let $Z(\tau_1),...,Z(\tau_n)$ be independent trials so that $Pr(Z(\tau_i)) = p_i$. Let $\bar{Z}(\tau_n) = \sum_{j=1}^{n} Z(\tau_j)$ and $\mu = \mathbb{E}(Z(\tau_n))$. Then with $k > 0$,*

$$Pr(|\bar{Z} - \mu| \geq k) \leq \frac{Var(Z)}{k^2}. \tag{10}$$

Proof. See Appendix B for details. □

Note that the Chebyshev inequality in Lemma IV.1 is defined in terms of the summation of random variables. Hence, the a sampling bound obtained directly from the Chebyshev inequality would conventionally be a bound on the cumulative error between all expected and sampled random variable values. This does not provide a bound on the error of our next sample. In adopting the iid requirement, the Bienaymé-Chebyshev inequality can provide such a bound, as shown in Lemma IV.2.

Lemma IV.2 (Bienaymé-Chebyshev Inequality). *Let $Z(\tau_1),...,Z(\tau_n)$ be independent and identically distributed trials so that $Pr(Z(\tau_i)) = p_i$. Let $\bar{Z}(\tau_n) = \frac{\sum_{j=1}^{n} Z(\tau_j)}{n}$ and $\mu = \mathbb{E}(\bar{Z}(\tau_n))$. Then with $k > 0$,*

$$Pr\left(\left|\bar{Z} - \mu\right| \geq k\right) \leq \frac{Var(Z)}{nk^2}. \tag{11}$$

Proof. See Appendix B for details. □

Although the Bienaymé-Chebyshev inequality in Lemma IV.2 can provide a sequential sampling bound, the identically distributed condition is quite limiting when the environment is not ergodic. The results obtained by the PIG algorithm, which learns the KL divergence ensemble average, reflects this shortcoming. To ensure a fair comparison, we show that PIG is equivalent to a Ped and use a sampling bound, analogous to our exposure bound, for PIG in Section IV-C.

B. Exposure Bound for Homogeneous Pep

We first will examine what Lemma IV.1 looks like when we insert our gamma distribution model on the homogeneous exposure rate of the Pep.

Lemma IV.3 (Chebyshev Inequality, $k = \lambda$). *Let $\Delta Z(\tau_1),...,\Delta Z(\tau_n)$ be independent Poisson exposure process trials so that $Pr(\Delta Z(\tau_i)) = p_i$. Let $\bar{Z}(\tau_n) = \sum_{j=1}^{n} \Delta Z(\tau_j)$ and $\lambda\beta = \mathbb{E}[Z(\tau_n)]$. Then,*

$$Pr\left(\left|\bar{Z} - \lambda\beta\right| \geq \lambda\right) \leq \frac{1}{\lambda}. \tag{12}$$

Proof. See Appendix B for details. □

In assuming heterogeneity of the Pep, we may extend Lemma IV.3 to provide a sequential-in-time bound as shown in Lemma IV.4, which is remarkably similar in form to Lemma IV.2.

Lemma IV.4 (Sequential-in-Time Inequality, $k = \lambda$). *Let $\Delta Z(\tau_1),...,\Delta Z(\tau_n)$ be independent Poisson exposure process increments so that $Pr(\Delta Z(\tau_i)) = p_i$. Let $\bar{Z}(\tau_n) = \frac{\sum_{j=1}^{n} \Delta Z(\tau_j)}{\tau_n}$ and $\lambda = \mathbb{E}[\bar{Z}(\tau_n)]$. Then,*

$$Pr\left(\left|\bar{Z} - \lambda\right| \geq \lambda\right) \leq \frac{1}{\lambda\beta}. \tag{13}$$

Proof. See Appendix B for details. □

We now consider a homogeneous Pep with $\Lambda(t, n) = \lambda t n$, where λ is the exposure per unit-time per sample, which provides us the following exposure inequality.

Theorem IV.5 (Exposure Inequality). *Let $\Delta Z(\tau_1),...,\Delta Z(\tau_n)$ be independent Poisson exposure process increments so that $Pr(\Delta Z(\tau_i)) = p_i$. Let $\bar{Z}(\tau_n) = \frac{\sum_{j=1}^{n} \Delta Z(\tau_j)}{n\tau_n}$ and $\lambda = \mathbb{E}[\bar{Z}(\tau_n)]$. Then,*

$$Pr\left(\left|\bar{Z} - \lambda\right| \geq \lambda^{\frac{3}{4}}\right) \leq \frac{1}{n\beta\sqrt{\lambda}}. \tag{14}$$

Proof. See Appendix B for details. □

While our exposure inequality in Theorem IV.5 provides an exposure bound for learning a single Pep, it does not yet provide a condition for the transition between uninformed and informed exploration across all bandit arms. Hence, we develop the Corollary IV.6.

Corollary IV.6 (Informed Policy Exposure Bound). *An uninformed-to-informed exploration algorithm has sufficiently explored the domain when*

$$\tau_{(b,n)} > \tau_{(b,1)} + \frac{1}{n_b c\sqrt{\lambda_b}}, \tag{15}$$

where $0 < c \leq 1$ and the inequality

$$\frac{1}{n_b \beta_b \sqrt{\lambda_b}} < c \tag{16}$$

holds true, where

$$b = \underset{i}{\operatorname{argmax}} \frac{1}{n_i \beta_i \sqrt{\lambda_i}}.$$

Proof. See Appendix B for details. □

Remark IV.7. *There are effectively no sampling guarantees until the bound from Theorem IV.5 is less than 1. Hence,* $0 \leq c < 1$.

The exposure bound in Theorem IV.5 extends the Chebyshev inequality such that a result analogous to the Bienaymé-Chebyshev inequality is achieved, but where samples need not be identically distributed. Moreover, guarantees on the Pep regression by Theorem IV.5 provide a principled condition for transitioning from uninformed exploration to informed exploration, which is in terms of the information exposure per unit time per sample.

C. Sample Bound for Baseline Method

Although we have an algorithm for uninformed-to-informed exploration, we need to examine a baseline uninformed-to-informed exploration algorithm alongside uninformed exploration algorithms, for completeness. As the PIG algorithm learns the average KL divergence, we may use the Ped to learn the average KL divergence and develop a sampling bound similar to Lemma IV.4 so as to facilitate a fair comparison between RAPTOR and PIG.

Lemma IV.8 (Sequential Sampling Inequality, $k = \lambda$)**.** *Let* $Z(\tau_1), ..., Z(\tau_n)$ *be independent Poisson exposure distribution trials so that* $Pr(Z(\tau_i)) = p_i$*. Let* $\bar{Z}(\tau_n) = \frac{\sum_{j=1}^{n} \Delta Z(\tau_j)}{n}$ *and* $\lambda = \mathbb{E}[\bar{Z}(\tau_n)]$*. Then,*

$$Pr\left(|\bar{Z} - \lambda| \geq \lambda\right) \leq \frac{1}{n\lambda}. \tag{17}$$

Proof. See Appendix B for details. □

Hence, we leverage a similar relationship between Theorem IV.5 and Corollary IV.6 to yield the following baseline uninformed-to-informed exploration policy.

Corollary IV.9 (Informed Policy Bound for PIG)**.** *An uninformed-to-informed exploration algorithm may be used for informed exploration over the entire state space once an agent has explored for a time* t *such that*

$$n > \frac{1}{\sqrt{\lambda_b}}, \tag{18}$$

where $0 < c \leq 1$ *and the inequality*

$$\frac{1}{n_b \lambda_b} < c \tag{19}$$

holds true, where

$$b = \underset{i}{\operatorname{argmax}} \frac{1}{n_i \lambda_i}.$$

Proof. See Appendix B for details. □

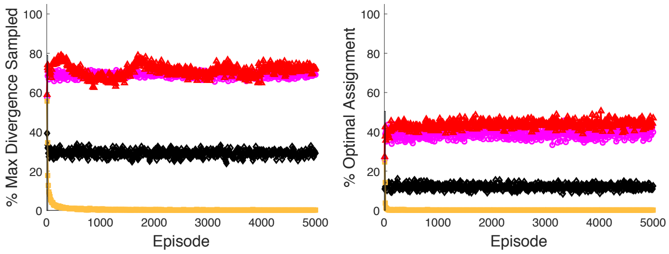
Using Corollary IV.9, we can now facilitate a fair comparison between the RAPTOR and PIG algorithms in Section V. Note that the PIG algorithm performs markedly better using Corollary IV.9 in Section V than without it in Section III.

V. Uninformed-to-Informed Results

Simulated and real-world data are used for experimentation as described in Sections V-A-V-B and V-C, respectively. In Sections V-A-V-B, the quantity $Z_i(\tau_{(i,n_i)}) = 0 \; \forall \; \tau_{(i,n_i)}$ is adopted immediately after visiting a state. This simplification of the adaptive prediction technique in RAPTOR is used to validate the Bayesian regression for the homogeneous Pep as well as the Pep-Cox Gaussian process approximation. In Section V-C, we set $Z_i(\tau_{i,n_i})$ equal to the last measured value of KL divergence at a bandit arm. In Section V-C we also test on a range of allowed plays per episode, $1 \leq \kappa \leq K$, on four real-world datasets.

A. Homogeneous Simulation

The number of bandit arms in this experiment is set to $K = 50$, and the initialization parameters for each method is ($\alpha_i = 0, \beta_i = 0$). For simplicity we use a heterogeneous set of seeded random variables for the homogeneous Poisson exposure rate and assume that the agent is able to play up to 6 bandit arms per episode; i.e., $\kappa = 6$. The results summarizing the performance of each approach in Figure 1 in the simulated heterogeneous set of homogeneous Pep reward-generating processes are viewable in Figures 3a and 3b.



(a) % Max Entropy Sampled (b) % Optimal Arm Play

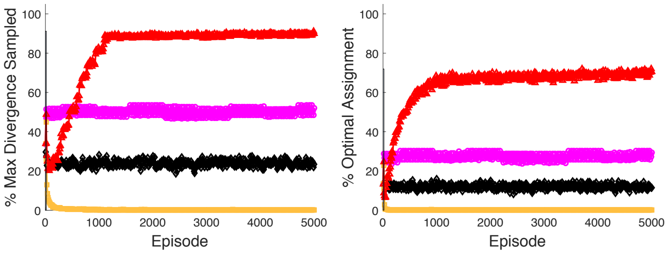
Figure 3: A simplified variant of RAPTOR explores the simulated data set, generated by a homogeneous Poisson exposure process, where 6 bandit arms are played per episode. Note that the high simulated noise in $\mathcal{Z}$ combined with several locations being similarly informative makes it such that a low accuracy in identifying the optimal action may still result in good entropy reduction performance.

B. Inhomogeneous Simulation

The conditions of the simulation in Section V-A are largely the same here; the parameters for each sensor location are again seeded randomly for the simulation, the plays per episode of the agent is the same ($\kappa = 6$), and the same priors are used as before ($\alpha_i = 0, \beta_i = 1$). However, instead of using a set of homogeneous Poisson exposure process, the observed rewards are generated by a Cox exposure process and are distorted by Gaussian noise as

$$\Lambda_i(\Delta t) = \left| \frac{\lambda_{r,i}}{8} (\sin(0.5 \Delta t \lambda_{r,i}) + 1) + \epsilon \right|, \tag{20}$$

where $\lambda_i(\Delta t)$ is the true underlying Poisson parameter, $\lambda_{r,i}$ is seeded randomly for the simulation and ϵ is Gaussian white noise; i.e., $\epsilon \sim N(0, \sigma^2)$. Note that it is reasonable to expect an accurate approximation of Gaussian data using the Poisson exposure parameter λ_i since it is convention to estimate large-parameter Poisson distributions with a Gaussian distribution.



(a) % Max Entropy Sampled (b) % Optimal Arm Play

Figure 4: A simplified variant of RAPTOR explores the simulated dataset, generated by a Cox exposure process, where 6 bandit arms may be played per episode.

C. Results on Real-World Datasets

As the results in Sections V-A and V-B are generated by artificial information metrics, we then apply RAPTOR on bandit problems using real-world datasets. For simplicity, each play at a bandit arm is modeled using a Gaussian distribution

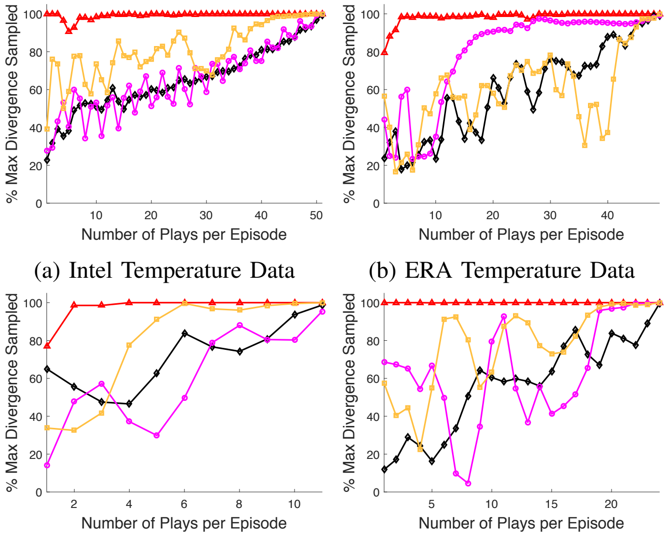
$$\mu_{\hat{q}} = \frac{\frac{\sigma_1^2}{M}\mu_{\hat{p}} + \sigma^2_{\hat{p},i}\bar{y}_i}{\frac{\sigma_1^2}{M} + \sigma^2_{\hat{p},i}} \quad \text{and} \quad \sigma_{\hat{q}}^{-2} = \left(\frac{\sigma_1^2}{M} + \sigma^2_{\hat{q}}\right)^{-1}, \tag{21}$$

where $\mu_{\hat{p}}$ is the prior mean, $\mu_{\hat{q}}$ is the posterior mean, $\sigma^2_{\hat{p},i}$ is the prior variance, and $\sigma^2_{\hat{q},i}$ is the posterior variance [10]. The KL divergence, D_{KL}, for scalar normal distributions (i.e., $d = 1$) is

$$\begin{aligned} D_{KL}(\hat{q}||\hat{p}) = 0.5[log\left(\frac{\sigma^2_{\hat{p}}}{\sigma^2_{\hat{q}}}\right) + tr\left[\left(\sigma^2{}_{\hat{p}}\right)^{-1}\sigma^2{}_{\hat{q}}\right] \\ - d + \left(\mu_{\hat{q}} - \mu_{\hat{p}}\right)^T \left(\sigma^2{}_{\hat{p}}\right)^{-1} \left(\mu_{\hat{q}} - \mu_{\hat{p}}\right)]. \end{aligned} \tag{22}$$

1) Intel Temperature Data: The Intel dataset contains the temperature measured (in Celsius) of the Intel lab at Berkeley (see Figure 6) between February 29th and April 5th of 2004 [11]. Due to highly variable quality of the recorded data, due to short or corrupted sensor feeds, we use approximately 5 days worth of data across 52 sensor feeds; i.e., $K = 52$. The RAPTOR algorithm is shown to outperform all baselines for this dataset in Figure 5a.

2) ERA Temperature Data: The European Research Area (ERA) temperature dataset contains the measured temperature at an altitude of 2 meters around the world (see Figure 7). We use data between January 1st 2011 to January 1st 2014 [12]. We limited our analysis to 50 randomly selected sensor feeds; i.e., $K = 50$.The RAPTOR algorithm is shown to outperform all baselines for this dataset in Figure 5b.



(a) Intel Temperature Data (b) ERA Temperature Data

(c) Ireland Windspeed Data (d) Washington Rain Data

Figure 5: Uninformed-to-informed exploration using RAPTOR is compared against baselines.

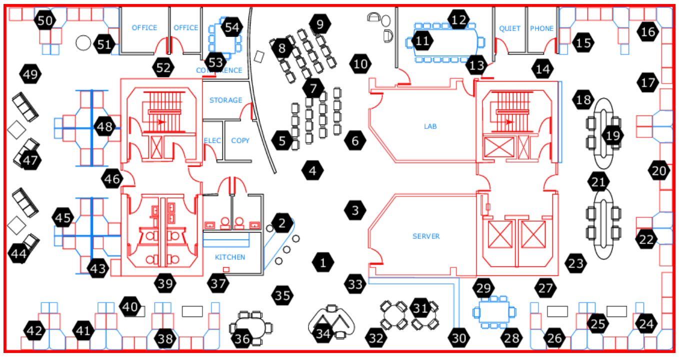


Figure 6: The Intel research lab has 54 sensors depicted, but 58 sensor feeds in their data file. Due to sensor quality constraints, we use 52 sensor feeds.

3) Ireland Windspeed Data: The Ireland windspeed dataset contains measured windspeed (in meters per second) at 12 stations across Ireland (see Figure 8a) between 1961-1978 [14]. We use all 12 sensing stations for experimentation; i.e., $K = 52$. The RAPTOR algorithm is shown to outperform all baselines for this dataset in Figure 5c.

4) Washington Rain Data: The Washington rainfall dataset (in millimeters) has many sensing locations contained in 272 gridded locations. Each resultant bandit arm for by the [15] data set is a weighted average over different precipitation measurements in adjacent gridded location, subject to the topology. The RAPTOR algorithm is shown to outperform all baselines for this dataset in Figure 5d.

D. Discussion of Results

In comparing the uninformed-to-informed exploration RAPTOR and PIG policies across the simulated and real-world datasets, we find that RAPTOR consistently outperforms PIG, as well as the uninformed exploration baselines. The primary weakness of average KL divergence estimation algorithms

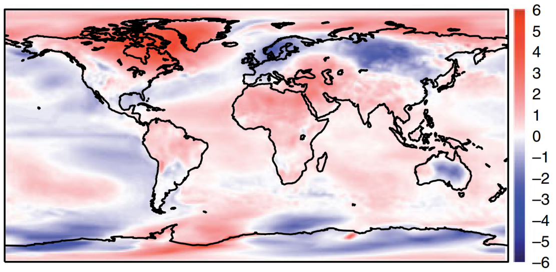
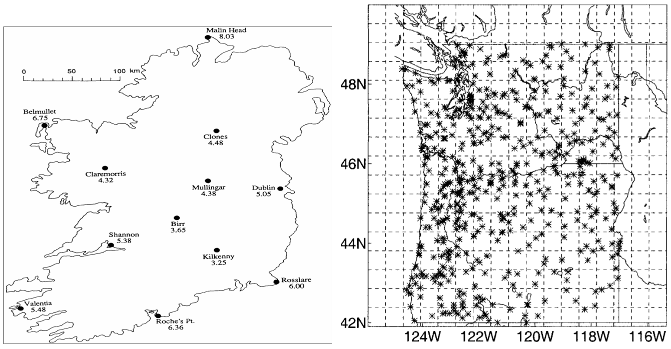


Figure 7: The magnitude, in Celsius, of anamolies in the European Research Area (ERA) interim dataset, relative to the means of the ERA-interim dataset for the year 2010 is shown [13].



(a) The Windspeed in Ireland was measured hourly at 12 sensing stations between the years 1961-1978 [14].

(b) The Washington rainfall dataset has many sensing locations contained in 272 gridded locations [15].

Figure 8: Maps of Ireland and Washington datasets.

such as PIG is that once the KL divergence at one bandit arm is underestimated, particularly if the KL divergence at the bandit arms is increasing, then the algorithms must sample all locations that have a higher expected value before the optimally informative arm will be sampled by the average KL divergence estimation algorithms.

By contrast, if the Lévy process estimation afforded by the Pep from Theorem IV.5 underestimates the KL divergence at a bandit arm, the information exposure rate $\Lambda(t,n)$ will cause that estimation to autonomously increase in a data-driven manner. Thus, the Pep may transition between sampling inefficiently informative bandit arms to the optimally informative bandit arm with less estimation inertia. The $Z(\tau_{(i,n_i)})$ term also adapts to the bounded approximation errors from Theorem IV.5 due to our approximation of the mutual information in (5) between $\mathcal{X}$ and $\mathcal{Y}$. This result is relatable to randomized policies, such as Information-Directed Search (IDS) [16], although the randomness in the measured KL divergence causes our deterministic uninformed-to-informed exploration policy to *adapt to randomness* in the environment, rather than having our agent make intentionally suboptimal sampling actions.

Although it would have been interesting to compare the IDS algorithm to RAPTOR, IDS requires knowledge on the probability of optimal action, which is difficult to know without a priori information. Furthermore, in the context of real-world datasets, actions are not necessarily known, but their ordering in the dataset is a deterministic mapping from past actions. In addition, it has been noted that human information-seeking behavior is highly correlated with the presence of KL divergence [17], which suggests that randomization may not be naturally selected for, in the context of information-driven human behaviors.

Lastly, in the context of the pure exploration problem, uninformed exploration methods such as ϵ-greedy [18] is equivalent to the random sampling baseline and Thompson sampling [19] is not usable as a reward function is not present. Although the explore-exploit tradeoffs of ϵ-greedy and Thompson sampling are of interest to future work, a crucial consideration of the explore-exploit dilemma is the efficacy of the exploration policies in their ability to acquire optimally informative samples.

VI. Conclusion

The benefit of informed exploration policies, which use information-theoretic quantities as the feedback signal, is that exploration policy may minimize the amount of undiscovered information, given sufficient knowledge about the information dynamics, in an environment. However, informed exploration policies have traditionally required preprocessing or a priori knowledge of the problem domain statistics, making them difficult to use in practice. Hence, we present an uninformed-to-informed exploration policy and a principled exposure bound for transitioning from uninformed exploration to an informed exploration. As a result, our algorithm outperforms baseline informed and uninformed exploration policies in real and synthetic datasets.

Appendix A
Synopsis

The following supplementary material covers proofs for completeness in Section B as well as comprehensive detail on experimental results in Section V.

Appendix B
Proofs

First, we briefly prove that the conjugate prior of the homogeneous Poisson exposure process (Pep) is a gamma distribution.

Fact B.1.
The gamma distribution is a conjugate prior of the homogeneous Poisson exposure process (Pep) such that

$$G\left(\lambda^{*} t|\alpha+z,\beta+t,z\right) \propto Pep\left(z|\lambda t\right) G\left(\lambda t|\alpha,\beta\right). \tag{23}$$

Proof. We apply the Poisson exposure process and gamma distribution to Bayes theorem as

$$f(\lambda^{*} t|\alpha^{*},\beta^{*},z) = C_{\lambda t}\frac{(\lambda t)^{z}e^{-\lambda t}}{\Gamma(z+1)}\cdot\frac{\beta^{\alpha}}{\Gamma(\alpha)}(\lambda t)^{\alpha-1}e^{-\beta\lambda t}. \tag{24}$$

In dropping the constant terms, we have

$$f(\lambda^{*} t|\alpha^{*},\beta^{*},z) \propto (\lambda t)^{z}e^{-\lambda t}\cdot(\lambda t)^{\alpha-1}e^{-\beta\lambda t}. \tag{25}$$

which resolves to

$$f(\lambda^* t|\alpha^*, \beta^*, z) \propto (\lambda t)^{\alpha+z-1} e^{-\lambda(\beta+t)}. \tag{26}$$

Therefore, given a Poisson exposure process with parameter λt, the posterior distribution will be proportional to the prior Gamma distribution with parameters, $\alpha^* = \alpha + z, \beta^* = \beta + t$. □

The Chebyshev and Bienaymé-Chebyshev inequalities in Lemmas IV.1 and IV.2 are provided verbatim [20], so we move on to providing proof for Lemma IV.3.

Lemma B.2 (Chebyshev Inequality, $k = \lambda$)**.** *Let $\Delta Z(\tau_1), ..., \Delta Z(\tau_n)$ be independent Poisson exposure process trials so that $Pr(\Delta Z(\tau_i)) = p_i$. Let $\bar{Z}(\tau_n) = \sum_{j=1}^{n} \Delta Z(\tau_j)$ and $\lambda\beta = \mathbb{E}[Z(\tau_n)]$. Then,*

$$Pr\left(|\bar{Z} - \lambda\beta| \geq \lambda\right) \leq \frac{1}{\lambda}. \tag{27}$$

Proof. From Lemma IV.1, we know that

$$Pr(|\bar{Z} - \mu| \geq k) \leq \frac{Var(Z)}{k^2}. \tag{28}$$

Inserting the mean and variance of our gamma distribution model on homogeneous Pep into (28) yields

$$Pr(|\bar{Z} - \frac{\alpha}{\beta}\beta| \geq k) \leq \frac{\frac{\alpha}{\beta^2}\beta}{k^2}, \tag{29}$$

which simplifies to

$$Pr(|\bar{Z} - \lambda\beta| \geq k) \leq \frac{\lambda}{k^2}. \tag{30}$$

In assigning, $\lambda = k$, we resolve the proof.

$$Pr(|\bar{Z} - \lambda\beta| \geq \lambda) \leq \frac{1}{\lambda}. \tag{31}$$

□

Similarly, the proof for Lemma IV.4 follows.

Lemma B.3 (Sequential-in-Time Inequality, $k = \lambda$)**.** *Let $\Delta Z(\tau_1), ..., \Delta Z(\tau_n)$ be independent Poisson exposure process increments so that $Pr(\Delta Z(\tau_i)) = p_i$. Let $\bar{Z}(\tau_n) = \frac{\sum_{j=1}^{n} \Delta Z(\tau_j)}{\tau_n}$ and $\lambda = \mathbb{E}[\bar{Z}(\tau_n)]$. Then,*

$$Pr\left(|\bar{Z} - \lambda| \geq \lambda\right) \leq \frac{1}{\lambda\beta}. \tag{32}$$

Proof. From Lemma IV.1, we know that

$$Pr(|\bar{Z} - \mu| \geq k) \leq \frac{Var(Z)}{k^2}. \tag{33}$$

Inserting the mean and variance of our gamma distribution model on homogeneous Pep into (33) yields

$$Pr(|\bar{Z} - \frac{\alpha}{\beta}| \geq k) \leq \frac{\frac{\alpha}{\beta^2}}{k^2}, \tag{34}$$

which simplifies to

$$Pr(|\bar{Z} - \lambda\beta| \geq k) \leq \frac{\lambda}{\beta k^2}. \tag{35}$$

In assigning, $\lambda = k$, we resolve the proof.

$$Pr(|\bar{Z} - \lambda\beta| \geq \lambda) \leq \frac{1}{\lambda\beta}. \tag{36}$$

□

Now, considering a Pep on the exposure rate per unit time per sample we obtain a proof for Theorem IV.5.

Theorem B.4 (Exposure Inequality)**.** *Let $\Delta Z(\tau_1), ..., \Delta Z(\tau_n)$ be independent Poisson exposure process increments so that $Pr(\Delta Z(\tau_i)) = p_i$. Let $\bar{Z}(\tau_n) = \frac{\sum_{j=1}^{n} \Delta Z(\tau_j)}{n\tau_n}$ and $\lambda = \mathbb{E}[\bar{Z}(\tau_n)]$. Then,*

$$Pr\left(|\bar{Z} - \lambda| \geq \lambda^{\frac{3}{4}}\right) \leq \frac{1}{n\beta\sqrt{\lambda}}. \tag{37}$$

Proof. From Lemma IV.1, we know that

$$Pr(|\bar{Z} - \mu| \geq k) \leq \frac{Var(Z)}{k^2}. \tag{38}$$

Inserting the mean and variance of our gamma distribution model on homogeneous Pep into (38) yields

$$Pr(|\bar{Z} - \lambda| \geq k) \leq \frac{\frac{\alpha}{(n\beta)^2}}{k^2}, \tag{39}$$

which simplifies to

$$Pr(|\bar{Z} - \lambda| \geq k) \leq \frac{\lambda}{n\beta k^2}. \tag{40}$$

In assigning, $\lambda^{\frac{3}{4}} = k$, we resolve the proof.

$$Pr(|\bar{Z} - \lambda\beta| \geq \lambda^{\frac{3}{4}}) \leq \frac{1}{n\beta\sqrt{\lambda}}. \tag{41}$$

□

Then Corollary IV.6 follows.

Corollary B.5 (Informed Policy Exposure Bound)**.** *An uninformed-to-informed exploration algorithm has sufficiently explored the domain when*

$$\tau_{(b,n)} > \tau_{(b,1)} + \frac{1}{n_b c\sqrt{\lambda_b}}, \tag{42}$$

where $0 < c \leq 1$ and the inequality

$$\frac{1}{n_b \beta_b \sqrt{\lambda_b}} < c \tag{43}$$

holds true, where

$$b = \underset{i}{\text{argmax}} \frac{1}{n_i \beta_i \sqrt{\lambda_i}}.$$

Proof. Given the result from Theorem IV.5, we know

$$Pr\left(|\bar{Z} - \lambda| \geq \lambda^{\frac{3}{4}}\right) \leq \frac{1}{n\beta\sqrt{\lambda}}. \tag{44}$$

When the right-hand side of (44) is between 1 and 0, we have meaningful guarantees on the error of our regression; i.e., when

$$c > \frac{1}{n\beta\sqrt{\lambda}}, \tag{45}$$

where $0 < c \leq 1$ and $\beta = \tau_n - \tau_1$. Then meaningful guarantees are available at a bandit arm when

$$\tau_n > \tau_1 + \frac{1}{nc\sqrt{\lambda}}. \tag{46}$$

Consequently, meaningful guarantees are available across the entire sensing domain once

$$\tau_{(b,n)} > \tau_{(b,1)} + \frac{1}{n_b c\sqrt{\lambda_b}}, \tag{47}$$

where

$$b = \operatorname*{argmax}_i \frac{1}{n_i \beta_i \sqrt{\lambda_i}}.$$

□

Lemma IV.8 provides similar guarantees to the baseline informed exploration, Predicted Information Gain (PIG).

Lemma B.6 (Sequential Sampling Inequality, $k = \lambda$)**.** *Let $Z(\tau_1), ..., Z(\tau_n)$ be independent Poisson exposure distribution trials so that $Pr(Z(\tau_i)) = p_i$. Let $\bar{Z}(\tau_n) = \frac{\sum_{j=1}^{n} \Delta Z(\tau_j)}{n}$ and $\lambda = \mathbb{E}[\bar{Z}(\tau_n)]$. Then,*

$$Pr\left(|\bar{Z} - \lambda| \geq \lambda\right) \leq \frac{1}{n\lambda}. \tag{48}$$

Proof. From Lemma IV.1, we know that

$$Pr(|\bar{Z} - \mu| \geq k) \leq \frac{Var(Z)}{k^2}. \tag{49}$$

Inserting the mean and variance of our gamma distribution model on homogeneous Pep into (49) yields

$$Pr(|\bar{Z} - \lambda| \geq k) \leq \frac{\frac{\alpha}{(n)^2}}{k^2}, \tag{50}$$

which simplifies to

$$Pr(|\bar{Z} - \lambda| \geq k) \leq \frac{\lambda}{nk^2}. \tag{51}$$

In assigning, $\lambda = k$, we resolve the proof.

$$Pr(|\bar{Z} - \lambda\beta| \geq \lambda) \leq \frac{1}{n\lambda}. \tag{52}$$

□

Lastly Corollary IV.9.

Corollary B.7 (Informed Policy Bound for PIG)**.** *An uninformed-to-informed exploration algorithm may be used for informed exploration over the entire state space once an agent has explored for a time t such that*

$$n > \frac{1}{\sqrt{\lambda_b}}, \tag{53}$$

where $0 < c \leq 1$ and the inequality

$$\frac{1}{n_b \lambda_b} < c \tag{54}$$

holds true, where

$$b = \operatorname*{argmax}_i \frac{1}{n_i \lambda_i}.$$

Proof. Given the result from Lemma IV.8, we know

$$Pr\left(|\bar{Z} - \lambda| \geq \lambda\right) \leq \frac{1}{n\lambda}. \tag{55}$$

When the right-hand side of (55) is between 1 and 0, we have meaningful guarantees on the error of our regression; i.e., when

$$c > \frac{1}{n\lambda}, \tag{56}$$

where $0 < c \leq 1$. Then meaningful guarantees are available at a bandit arm when

$$n > \frac{1}{c\lambda}. \tag{57}$$

Consequently, meaningful guarantees are available across the entire sensing domain once

$$n_b > \frac{1}{c\lambda_b}, \tag{58}$$

where

$$b = \operatorname*{argmax}_i \frac{1}{n_i \lambda_i}.$$

□

Acknowledgments

This work is sponsored by the Department of Energy Award Number DE-FE0012173, the Air Force Office of Scientific Research Award Number FA9550-14-1-0399, and the Air Force Office of Scientific Research Young Investigator Program Number FA9550-15-1-0146.

References

[1] F. Creutzig, A. Globerson, and N. Tishby, "Past-future information bottleneck in dynamical systems," *Physical Review E*, vol. 79, no. 4, p. 041925, 2009.

[2] A. Hobson, "A new theorem of information theory," *Journal of Statistical Physics*, vol. 1, no. 3, pp. 383–391, 1969.

[3] D. Y. Little and F. T. Sommer, "Learning and exploration in action-perception loops," *Frontiers in neural circuits*, vol. 7, 2013.

[4] T. Kim, A. V. Nefian, and M. J. Broxton, "Photometric recovery of ortho-images derived from apollo 15 metric camera imagery," in *Advances in Visual Computing*. Springer, 2009, pp. 700–709.

[5] ——, "Photometric recovery of apollo metric imagery with lunar-lambertian reflectance," *Electronics letters*, vol. 46, no. 9, pp. 631–633, 2010.

[6] J. Møller, A. R. Syversveen, and R. P. Waagepetersen, "Log gaussian cox processes," *Scandinavian journal of statistics*, vol. 25, no. 3, pp. 451–482, 1998.

[7] Z. Zhou, D. S. Matteson, D. B. Woodard, S. G. Henderson, and A. C. Micheas, "A spatio-temporal point process model for ambulance demand," *arXiv preprint arXiv:1401.5547*, 2014.

[8] R. P. Adams, I. Murray, and D. J. MacKay, "Tractable nonparametric bayesian inference in poisson processes with gaussian process intensities," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 9–16.

[9] T. Gunter, C. Lloyd, M. A. Osborne, and S. J. Roberts, "Efficient bayesian nonparametric modelling of structured point processes," *arXiv preprint arXiv:1407.6949*, 2014.

[10] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*, 3rd ed. CRC press, 2013.

[11] P. Bodik, W. Hong, C. Guestrin, S. Madden, M. Paskin, and R. Thibaux, "Intel lab data," Intel Berkely Research Lab, Tech. Rep., Feb 2004. [Online]. Available: http://db.csail.mit.edu/labdata/labdata.html

[12] P. Berrisford, D. Dee, K. Fielding, M. Fuentes, P. Kallberg, S. Kobayashi, and S. Uppala, "The era-interim archive." 2009.

[13] D. Dee, S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer *et al.*, "The era-interim reanalysis: Configuration and performance of the data assimilation system," *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 656, pp. 553–597, 2011.

[14] J. Haslett and A. E. Raftery, "Ireland wind data set," Trinity College and University of Washington, Tech. Rep., 1961-1978. [Online]. Available: http://lib.stat.cmu.edu/datasets/wind.desc

[15] M. Widmann and C. S. Bretherton, "Validation of mesoscale precipitation in the ncep reanalysis using a new gridcell dataset for the northwestern united states," *Journal of Climate*, vol. 13, no. 11, pp. 1936–1950, 2000.

[16] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," in *Advances in Neural Information Processing Systems*, 2014, pp. 1583–1591.

[17] L. Itti and P. F. Baldi, “Bayesian surprise attracts human attention,” in *Advances in neural information processing systems*, 2005, pp. 547–554.

[18] M. Tokic, “Adaptive ε-greedy exploration in reinforcement learning based on value differences,” in *KI 2010: Advances in Artificial Intelligence*. Springer, 2010, pp. 203–210.

[19] A. Gopalan, S. Mannor, and Y. Mansour, “Thompson sampling for complex online problems,” in *Proceedings of The 31st International Conference on Machine Learning*, 2014, pp. 100–108.

[20] C. Heyde and E. Seneta, “Studies in the history of probability and statistics. xxxi. the simple branching process, a turning point test and a fundamental inequality: A historical note on ij bienaymé,” *Biometrika*, vol. 59, no. 3, pp. 680–683, 1972.

PLACE
PHOTO
HERE

Allan Axelrod Biography text here.

PLACE
PHOTO
HERE

Dr. Girish Chowdhary Biography text here.