Neural Inverse CDF Sampling

Andrew Stirn: andrew.stirn@columbia.edu
Tony Jebara: jebara@cs.columbia.edu

Problem

Thompson sampling (TS), a strategy to balance *exploration* and *exploitation* in multi-armed bandit problems, enjoys theoretical guarantees and strong empirical performance in a variety of domains. However, TS assumes causal relationships between selected actions and received rewards.

Contributions

This article relaxes those assumptions by considering the *noncompliance* setting where a chosen action is not necessarily the implemented action. We formalize TS for both *observed noncompliance* and *latent noncompliance*. Further, we empirically demonstrate the performance benefit of our approach for a variety of environments with differing levels of stochastic noncompliance.

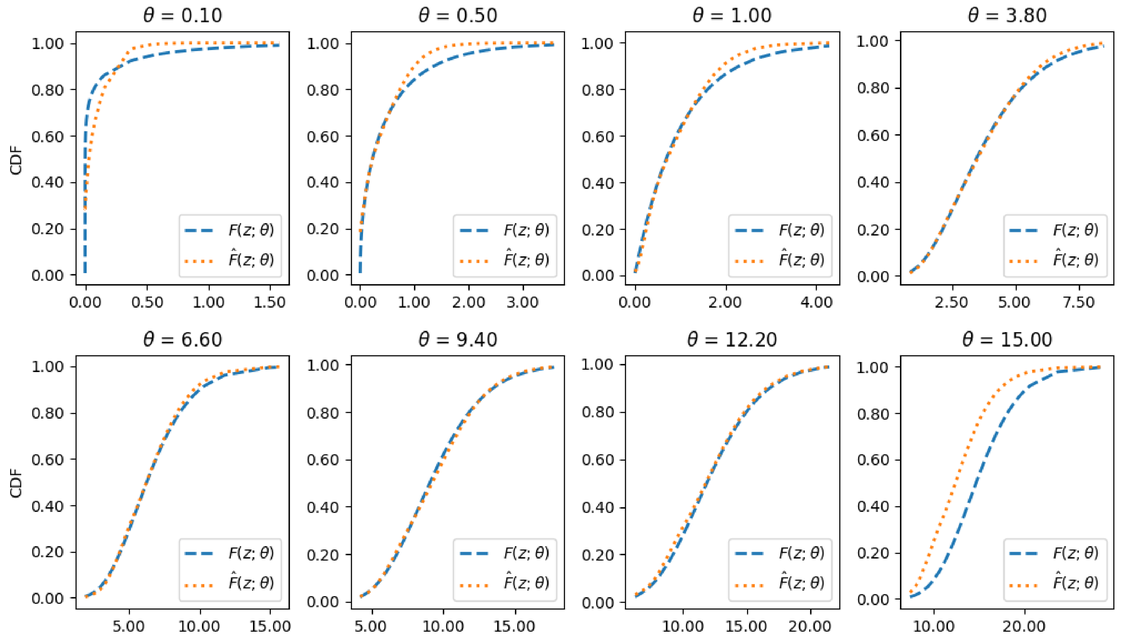
Inverse CDF

TS applies Bayesian inference in a sequential manner. As data is collected, the posterior transitions from a high entropy (explorative) prior to a more peaked (exploitative) distribution. To apply TS to a Bandit model, one needs:

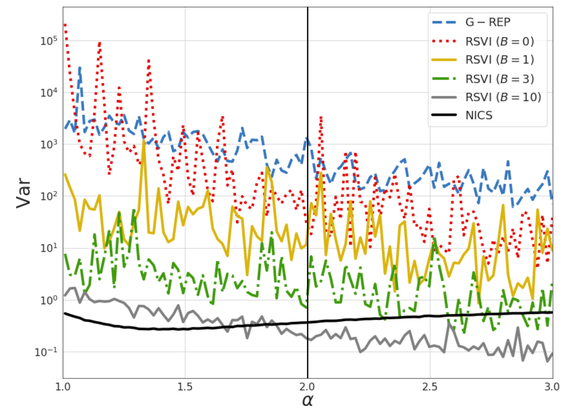
Invertible INN

TS applies Bayesian inference in a sequential manner. As data is collected, the posterior transitions from a high entropy (explorative) prior to a more peaked (exploitative) distribution. To apply TS to a Bandit model, one needs:

CDF Approximation for Gamma Distribution



Reparameterize



Results

	Perplexity Scores	
	Softmax	NGS
NVLDA	8.95	7.04
ProDLDA	9.55	6.90

Simulated Environments

BLAH

References & Code

[1] Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. In *Biometrika 1933*.

Source code available at:
`github.com/astirn/ExpandedThompsonSampling/`