From Prediction to Decisions: The importance of Joint predictive distribution

Yingru Li
yingruli@link.cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen, China

September 20, 2022

Motivations

- The Neural Testbed: Evaluating Joint Predictions [Osband et al., 2021]
- From Predictions to Decisions: The Importance of Joint Predictive Distributions [Wen et al., 2021]

Outline

Part I: Importance of Joint prediction for Decision making

- Combinatorial decision problems in recommendation systems
- Sequential decision problem
- Multi-armed bandits

Part II: Empirical evaluation of joint prediction and its correlation to decision making

Discussion

Data sequence

- Consider a sequence of pairs $((X_t, Y_{t+1}) : t = 0, 1, 2, \ldots)$;

$$\left(\underbrace{X_t}_{\text{feature vector}} \overset{i.i.d}{\sim} P_X \;,\; \underbrace{Y_{t+1}}_{\text{target label}} \right)$$

- The conditional distribution $\mathcal{E}$ is referred to as the environment.
- The environment $\mathcal{E}$ is random; and this reflects the agent's uncertainty about how labels are generated given features.
- Each target label $Y_{t+1} \perp\!\!\!\perp$ all other data $| \, X_t$ and

$$\mathbb{P}\left(Y_{t+1} \in \cdot \mid \mathcal{E}, X_t\right) = \mathcal{E}(\cdot \mid X_t)$$

And we have $\mathbb{P}\left(Y_{t+1} \in \cdot \mid X_t\right) = \mathbb{E}\left[\mathcal{E}\left(\cdot \mid X_t\right) \mid X_t\right]$.

Supervised learning

- Supervised learning: an agent that learns about the environment $\mathcal{E}$ from a training dataset

$$\mathcal{D}_T \equiv \left((X_t, Y_{t+1}) : t = 0, 1, \ldots, T-1 \right),$$

and aims to predict the target labels

$$Y_{T+1:T+\tau} \equiv (Y_{T+1}, \ldots, Y_{T+\tau})$$

at τ feature vectors $X_{T:T+\tau-1} \equiv (X_T, \ldots, X_{T+\tau-1})$.

Predictive distribution

- Conditioned on the environment $\mathcal{E}$, a predictive distribution over the target labels is given by

$$P^*_{T+1:T+\tau} \equiv \mathbb{P}\left(Y_{T+1:T+\tau} \in \cdot \mid \mathcal{E}, X_{T:T+\tau-1}\right).$$

- Conditioned instead on the training data, the predictive distribution becomes

$$\begin{aligned}\bar{P}_{T+1:T+\tau} &\equiv \mathbb{P}\left(Y_{T+1:T+\tau} \in \cdot \mid \mathcal{D}_T, X_{T:T+\tau-1}\right) \\ &= \mathbb{E}\left[\mathcal{E}(Y_{T+1:T+\tau} \in \cdot \mid X_{T:T+\tau-1}) \mid \mathcal{D}_T, X_{T:T+\tau-1}\right] \\ &= \mathbb{E}\left[\prod_{t=T}^{T+\tau-1} \mathcal{E}(Y_{t+1} \in \cdot \mid X_t) \mid \mathcal{D}_T, X_{T:T+\tau-1}\right]\end{aligned}$$

- Since $\mathcal{E}$ is random, the conditional expectation $\mathbb{E}\left[\mathcal{E}(\cdot) \mid \mathcal{D}_T\right]$ denotes the true posterior of $\mathcal{E}$ given $\mathcal{D}_T$.
- $\bar{P}_{T+1:T+\tau}$ represents the result of perfect (Bayesian posterior) inference.

Problems of perfect inference for predictive distribution

- Problem 1 (Computational tractability):
  - Perfect inference is computationally tractable if conjugate property exists for the environment $\mathcal{E}$, e.g. linear Gaussian, Beta-Bernoulli, and some GPs.
  - Perfect inference is usually computationally intractable for the environments of interest (e.g. Nonlinear models or Neural networks).
- Problem 2 (Computational efficiency):
  - For linear Gaussian model, posterior update (perfect inference) can be computed using rank-one update rule.
  - For GPs, the computational complexity of posterior update (perfect inference) is dominated by $\mathcal{O}(N^3)$ where N is the number of data.
- To tackle these issues, consider agents that perform approximate inference.

Approximate predictive distribution

- Consider agents that represent the approximation in terms of a generative model.
- The agent's predictions are parameterized by a vector θ_T that the agent (only) learns from the training data $\mathcal{D}_T$.
- The vector θ_T is conditionally independent of $\mathcal{E}$ conditioned on $\mathcal{D}_T$.

$$\theta_T \perp\!\!\!\perp \mathcal{E} \mid \mathcal{D}_T$$

- For any inputs $X_{T:T+\tau-1}, \theta_T$ determines a predictive distribution, which could be used to sample imagined outcomes $\hat{Y}_{T+1:T+\tau}$.
- Hence, the agent's τ^{th} -order predictive distribution is given by

$$\hat{P}_{T+1:T+\tau} \equiv \mathbb{P}\left(\hat{Y}_{T+1:T+\tau} \in \cdot \mid \theta_T, X_{T:T+\tau-1}\right)$$

Approximate predictive distribution

- Consider agents that represent the approximation in terms of a generative model.
- The agent's predictions are parameterized by a vector θ_T that the agent (only) learns from the training data $\mathcal{D}_T$.
- The vector θ_T is conditionally independent of $\mathcal{E}$ conditioned on $\mathcal{D}_T$.

$$\theta_T \perp\!\!\!\perp \mathcal{E} \mid \mathcal{D}_T$$

- For any inputs $X_{T:T+\tau-1}, \theta_T$ determines a predictive distribution, which could be used to sample imagined outcomes $\hat{Y}_{T+1:T+\tau}$.
- Hence, the agent's τ^{th} -order predictive distribution is given by

$$\hat{P}_{T+1:T+\tau} \equiv \mathbb{P}\left(\hat{Y}_{T+1:T+\tau} \in \cdot \mid \theta_T, X_{T:T+\tau-1}\right)$$

Marginal vs. joint predictive distributions

- When $\tau = 1$, we alternatively use $\hat{P}_{T+1}, \bar{P}_{T+1}$, and P^*_{T+1} to denote $\hat{P}_{T+1:T+\tau}, \bar{P}_{T+1:T+\tau}$, and $P^*_{T+1:T+\tau}$, respectively.
- Marginal prediction: $\tau = 1, \hat{P}_{T+1}$ predicts the label Y_{T+1} for a single input X_T.
- Joint prediction: $\tau > 1, \hat{P}_{T+1:T+\tau}$ represents a joint prediction over labels at τ input features.

Marginal vs. joint predictive distributions: Coin flipping example

- $(Y_{t+1} : t = 0, 1, \ldots)$: repeated tosses of a possibly biased coin with unknown probability p of heads, with $Y_{t+1} = 1$ and $Y_{t+1} = 0$ indicating heads and tails, respectively.
- Consider two agents with different beliefs:
  - Agent 1 assumes $p = 2/3$ and models the outcome of each coin toss as independent conditioned on p.
  - Agent 2 assumes that $p = 1$ with probability $2/3$ and $p = 0$ with probability $1/3$; that is, the coin either produces only heads or only tails.
- Let $\hat{Y}^1_{t+1}$ and $\hat{Y}^2_{t+1}$ denote the outcomes imagined by the two agents.
- Despite their differing assumptions, the two agents generate identical marginal predictive distributions:

$$\mathbb{P}\left(\hat{Y}^1_{t+1} = 0\right) = \mathbb{P}\left(\hat{Y}^2_{t+1} = 0\right) = 1/3$$

Marginal vs. joint predictive distributions: Coin flipping example

- Identical marginal predictive distributions:

$$\mathbb{P}\left(\hat{Y}^1_{t+1} = 0\right) = \mathbb{P}\left(\hat{Y}^2_{t+1} = 0\right) = 1/3$$

- Joint predictions of these two agents differ for $\tau > 1$:

$$\mathbb{P}\left(\hat{Y}^1_1, \ldots, \hat{Y}^1_\tau = 0\right) = 1/3^\tau < 1/3 = \mathbb{P}\left(\hat{Y}^2_1, \ldots, \hat{Y}^2_\tau = 0\right)$$

- Evaluating marginal predictions cannot distinguish between the two agents, though for a specific prior distribution over p, one agent could be right and the other wrong.
- Conclusion: One **must evaluate joint predictions** to make this distinction.

Cross-entropy loss for evaluating marginal and joint predictions

- Cross-entropy loss to evaluate marginal predictive distributions.

$$\mathbf{d}_{\mathrm{CE}}^{1} \equiv -\mathbb{E}\left[\log \hat{P}_{T+1}\left(Y_{T+1}\right)\right]$$

  where the expectation is over both $\hat{P}_{T+1}$ and Y_{T+1}.
- the superscript " 1 " in $\mathbf{d}_{\mathrm{CE}}^{1}$ indicates that this evaluates marginal predictions.
- Note that the marginal distribution $\hat{P}_{T+1}$ is random because it depends on θ_T and X_T.

Cross-entropy loss for evaluating marginal and joint predictions

- Straightforward to extend the cross-entropy loss to assess joint predictive distributions.
- For any $\tau = 1, 2, \ldots$, we define the τ^{th} -order crossentropy loss:

$$\mathbf{d}_{\text{CE}}^{\tau} \equiv -\mathbb{E}\left[\log \hat{P}_{T+1:T+\tau}\left(Y_{T+1:T+\tau}\right)\right]$$

  where the expectation is over $\hat{P}_{T+1:T+\tau}$ and $Y_{T+1:T+\tau}$.
- Note that the τ^{th} -order joint distribution $\hat{P}_{T+1:T+\tau}$ is also random, since it depends on θ_T and $X_{T:T+\tau-1}$.

Kullbeck-Leibler divergence

- For a more elegant mathematical analysis, it can be helpful to offset the metric by a baseline to convert it into the Kullback-Leibler (KL) divergence.
- The τ^{th}-order expected KL-divergence with respect to $\bar{P}$ is defined by

$$\mathbf{d}_{\mathrm{KL}}^{\tau} \equiv \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\bar{P}_{T+1:T+\tau} \| \hat{P}_{T+1:T+\tau}\right)\right]$$

  where the expectation is over the distributions $\bar{P}_{T+1:T+\tau}$ and $\hat{P}_{T+1:T+\tau}$, which depend in turn on the data $\mathcal{D}_T$, the agent parameters θ_T, and the τ inputs $X_{T:T+\tau-1}$.
- Note that KL-divergence is minimized when $\hat{P}_{T+1:T+\tau} = \bar{P}_{T+1:T+\tau}$, with the minimum being zero.

Relation between Cross-Entropy and Kullbeck-Leibler divergence

- Further, the two metrics are related according to

$$\mathbf{d}_{\mathrm{KL}}^{\tau} = \mathbf{d}_{\mathrm{CE}}^{\tau} + \mathbb{E}\left[\log \bar{P}_{T+1:T+\tau}\left(Y_{T+1:T+\tau}\right)\right].$$

- Since $\bar{P}_{T+1:T+\tau}$ does not depend on the agent, our measure of KL-divergence and the cross-entropy loss are effectively equivalent in the sense that they only differ by a constant that does not depend on the agent.
- Since $\bar{P}_{T+1:T+\tau}\left(Y_{T+1:T+\tau}\right)$ does not depend on the agent, our two metrics rank agents identically.

Evaluation on Cross-Entropy and Kullbeck-Leibler divergence

- An unbiased estimate of cross-entropy loss can be computed based on a test data sample, according to
$$\mathbf{d}_{\text{CE}}^{\tau} \approx -\log \hat{P}_{T+1:T+\tau}\left(Y_{T+1:T+\tau}\right)$$
- The same is not true for $\mathbf{d}_{\text{KL}}^{\tau}$, which can only be estimated if also given an estimate of $\mathbb{E}\left[\log \bar{P}_{T+1:T+\tau}\left(Y_{T+1:T+\tau}\right)\right]$.

Conclusion on the Metrics

- Hence, $\mathbf{d}^{\tau}_{\mathrm{KL}}$ serves only as conceptual tools in our analysis and not an evaluation metric that can be applied with empirical data.
- While it ranks agents identically with $\mathbf{d}^{\tau}_{\mathrm{CE}}$, $\mathbf{d}^{\tau}_{\mathrm{KL}}$ is more natural as a metric since its minimum is zero and it accommodates more elegant analysis.

Error in predictions versus environment

- Our $\mathbf{d}_{\mathrm{KL}}^{\tau}$ metric assesses error incurred by the predictive distribution $\hat{P}_{T+1:T+\tau}$.
- A common approach to generating such a predictive distribution:
    1. estimating a posterior distribution over environments
    2. using that posterior distribution to generate the predictive distribution.
- In such a context, θ_T parameterizes the estimated posterior distribution.
- Let $\hat{\mathcal{E}}$ be an imaginary environment sampled from this posterior distribution.

Error in predictions versus environment

- To offer some intuition for $\mathbf{d}_{\mathrm{KL}}^{\tau}$, we consider in this section its relation to the KL-divergence between the distributions of the true and imaginary environments.
- Let $\hat{Y}_{T+1:T+\tau}$ denote a sequence of imaginary outcomes, with each $\hat{Y}_{t+1}$ sampled independently from $\hat{\mathcal{E}}\left(\cdot \mid X_t\right)$.
- If the support of the input distribution P_X is exhaustive, the support of the imaginary environment distribution $\mathbb{P}\left(\hat{\mathcal{E}} \in \cdot \mid \theta_T\right)$ contains that of the true environment distribution $\mathbb{P}(\mathcal{E} \in \cdot \mid \mathcal{D}_T)$, and the environment distributions satisfy suitable regularity conditions, then

$$\lim_{\tau \to \infty} \mathbf{d}_{\mathrm{KL}}^{\tau} = \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(\mathcal{E} \in \cdot \mid \mathcal{D}_T\right) \| \mathbb{P}\left(\hat{\mathcal{E}} \in \cdot \mid \theta_T\right)\right)\right]$$

Why use $\mathbf{d}_{\mathrm{KL}}^{\tau}$ instead of KL between the true and imaginary envs?

$$\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(\mathcal{E} \in \cdot \mid \mathcal{D}_T\right) \| \mathbb{P}\left(\hat{\mathcal{E}} \in \cdot \mid \theta_T\right)\right)\right] \tag{1}$$

- Practical agent design often do not satisfy the requisite regularity conditions and hence eq. (1) becomes infinite
  - For example, it is common to approximate the posterior distribution $\mathcal{E}$ using an ensemble of environment models (see, e.g., Lu & Van Roy (2017)). Such an ensemble represents a distribution with finite support though the posterior may have infinite support.
  - On the other hand, for any finite $\tau, \mathbf{d}_{\mathrm{KL}}^{\tau}$ is finite.
- Second, $\mathbf{d}_{\mathrm{CE}}^{\tau}$, which is equivalent to $\mathbf{d}_{\mathrm{KL}}^{\tau}$ up to a constant, can be computed based on data, whereas computing eq. (1) requires access to the posterior distribution of $\mathcal{E}$.
- Finally, as we will establish later, $\mathbf{d}_{\mathrm{KL}}^{\tau}$ with finite τ is sufficient to support effective decisions in downstream tasks such as multi-armed bandits.

Universality of $\mathbf{d}_{KL}^{\tau}$

- For any τ, accuracy in terms of $\mathbf{d}_{\mathrm{KL}}^{\tau}$ is sufficient to guarantee an effective decision if the decision is judged in relation only to $Y_{T+1:T+\tau}$.
- In particular, suppose an action a selected from a set $\mathcal{A}$ results in an expected reward

$$\mathbb{E}\left[r\left(a, Y_{T+1:T+\tau}\right) \mid \mathcal{D}_T, X_{T:T+\tau-1}\right]$$
$$= \sum_{y_{T+1:T+\tau}} \bar{P}_{T+1:T+\tau}\left(y_{T+1:T+\tau}\right) r\left(a, y_{T+1:T+\tau}\right),$$

where r is a reward function with range $[0,1]$.

Universality of $\mathbf{d}_{KL}^{\tau}$

The following result bounds the loss in expected reward of a decision that is based on the estimate $\hat{P}_{T+1:T+\tau}$ instead of the posterior $\bar{P}_{T+1:T+\tau}$

Proposition 1.

If an action $\hat{a} \in \mathcal{A}$ maximizes

$$\sum_{y_{T+1:T+\tau}} \hat{P}_{T+1:T+\tau}\left(y_{T+1:T+\tau}\right) r\left(a, y_{T+1:T+\tau}\right)$$

then

$$\mathbb{E}\left[r\left(\hat{a}, Y_{T+1:T+\tau}\right)\right] \geqslant \max_{a \in \mathcal{A}} \mathbb{E}\left[r\left(a, Y_{T+1:T+\tau}\right)\right] - \sqrt{2\mathbf{d}_{\mathrm{KL}}^{\tau}}$$

In this sense, $\mathbf{d}_{\mathrm{KL}}^{\tau}$ is a universal evaluation metric: its value ensures a level of performance in any decision problem.

Outline

Problem setup

- Consider the problem of a customer interacting with a recommendation system that proposes a selection of $K > 1$ movies from an inventory of N movies $X_1, \ldots, X_N$.
- Each $X_i \in \mathbb{R}^d$ describes the features of movie i, and d is the feature dimension.
- We model the probability that a user will enjoy movie i by a logistic model $Y_i \sim \text{logit}\left(\phi_*^T X_i\right)$, where logit is the standard logistic function.
- Note that $\phi_* \in \mathbb{R}^d$ describes the preferences of the user, which is not fully known to the recommendation system and can be viewed as a random variable.
- Goal: maximize the probability that the user enjoys at least one of the $K > 1$ recommended movies.

Concrete example

- The user ϕ_* is drawn from two possible user types $\{\phi_1, \phi_2\}$
- Recommendation propose $K = 2$ movies from an inventory $\{X_1, X_2, X_3, X_4\}$
- These values are chosen to set up a tension between optimization over marginal (each X_i individually) and joint (pairs of X_i, X_j) predictions.

	$X_1 = (10, -10)$	$X_2 = (-10, 10)$	$X_3 = (1, 0)$	$X_4 = (0, 1)$
$\phi_1 = (1, 0)$	1	0	0.73	0.5
$\phi_2 = (0, 1)$	0	1	0.5	0.73
$\phi \sim \text{Unif}(\phi_1, \phi_2)$	0.5	0.5	0.62	0.62

Table: Expected probability to watch a movie under different user features, correct to two decimal places.

Concrete example

	$X_1 = (10, -10)$	$X_2 = (-10, 10)$	$X_3 = (1, 0)$	$X_4 = (0, 1)$
$\phi_1 = (1, 0)$	1	0	0.73	0.5
$\phi_2 = (0, 1)$	0	1	0.5	0.73
$\phi \sim \text{Unif}(\phi_1, \phi_2)$	0.5	0.5	0.62	0.62

Table: Expected probability to watch a movie under different user features, correct to two decimal places.

- An agent that optimizes the expected probability for each movie individually will end up recommending the pair (X_3, X_4) to an unknown $\phi \sim \text{Unif}(\phi_1, \phi_2)$.
- An agent considers the joint predictive distribution for $\tau \geqslant K = 2$ can see that instead selecting the pair (X_1, X_2).

Outline

Problem setup

- Data pair (X_t, Y_{t+1}) arrives sequentially, one at a time.
- At each time t, the agent needs to compute parameters θ_t based on previously observed data pairs $\mathcal{D}_t = (X_0, Y_1, X_1, \ldots, X_{t-1}, Y_t)$.
- Then, a new data pair (X_t, Y_{t+1}) arrives. We assume that the feature vector X_t 's are unconditionally independent, but not necessarily identically distributed.
- The target label Y_{t+1} is conditionally independently sampled from the distribution $\mathcal{E}(\cdot \mid X_t)$, where $\mathcal{E}$ is the environment.

Problem setup

- The agent's objective is to minimize the expected cumulative KL-divergence in the first T time steps:

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\bar{P}_{t+1} \| \hat{P}_{t+1}\right)\right],$$

where

$$\bar{P}_{t+1} = \mathbb{P}\left(Y_{t+1} \in \cdot \mid \mathcal{D}_t, X_t\right)$$
$$\hat{P}_{t+1} = \mathbb{P}\left(\hat{Y}_{t+1} \in \cdot \mid \theta_t, X_t\right)$$

for all time t. Note that this cumulative KL-divergence (5) only depends on the marginal distributions $\bar{P}_{t+1}$ and $\hat{P}_{t+1}$.

- Also note that this performance metric is 0 if the agent predicts the exact posterior at each time t.

Incremental update

- We consider a setting where an agent needs to incrementally update its parameters as data arrive.
- Specifically, at time $t = 0$, the agent chooses its parameters θ_0 based on its prior knowledge; and then at each time $t = 0, 1, \ldots$, the agent updates its parameters incrementally by sampling from a distribution that only depends on $\theta_t, (X_t, Y_{t+1})$, and t:

$$\theta_{t+1} \sim \mathbb{P}\left(\theta_{t+1} \in \cdot \mid \theta_t, X_t, Y_{t+1}, t\right). \tag{2}$$

- In other words, conditioning on $\left(\theta_t, X_t, Y_{t+1}\right), \theta_{t+1}$ is independent of the dataset $\mathcal{D}_t$ and the environment $\mathcal{E}$.

Remark on the incremental update

- Note that the incremental update rule in eq. (2) is general: in particular, $\mathcal{D}_t$ could itself be recorded in θ_t. This would allow θ_{t+1} to depend on $\mathcal{D}_t$ in an arbitrary manner. However, such an approach can be impractical when there is a high volume of data.
- In particular, one may want to avoid sifting through a growing $\mathcal{D}_t$ at each time step.
- In many practical applications, it is desirable for the agent to update θ_{t+1} with fixed memory space and fixed per-step computational complexity, such as the standard SGD (Goodfellow et al., 2016) and Adam (Kingma & Ba, 2015) algorithms do.

Theorem for sequenctial prediction problem

Theorem 1.

For an agent with incremental update eq. (2), *for any time* $t = 0, 1, \ldots, T-1$ *and any* $\epsilon \geqslant 0$, *if*

$$\sum_{t'=t}^{T-1} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\bar{P}_{t'+1} \| \hat{P}_{t'+1}\right)\right] \leqslant \epsilon,$$

then we have

$$\mathbb{I}\left(Y_{t+1:T}; \theta_t \mid X_{t:T-1}\right) \geqslant \mathbb{I}\left(Y_{t+1:T}; \mathcal{D}_t \mid X_{t:T-1}\right) - \epsilon.$$

Remark on the theorem

- Notice that ϵ measures the performance loss of the agent; $\mathbb{I}\left(Y_{t+1:T}; \mathcal{D}_t \mid X_{t:T-1}\right)$ is the conditional information in $\mathcal{D}_t$ about the joint distribution of $Y_{t+1:T}$; and similarly $\mathbb{I}\left(Y_{t+1:T}; \theta_t \mid X_{t:T-1}\right)$ is the conditional information about $Y_{t+1:T}$ retained in θ_t.
- Also notice that

$$\mathbb{I}\left(Y_{t+1:T}; \mathcal{D}_t \mid X_{t:T-1}\right) \geqslant \mathbb{I}\left(Y_{t+1:T}; \theta_t \mid X_{t:T-1}\right)$$

  always holds due to data processing inequality.
- In other words, Theorem 4.1 states that to be ϵ-near-optimal, an agent with incremental update must retain in θ_t all information in $\mathcal{D}_t$ about the joint distribution of $Y_{t+1:T}$, except ϵ nats
- We conjecture that results similar to Theorem 4.1 also hold in broader classes of sequential decision problems, such as multi-armed bandit problems discussed in Section 5, but we leave the formal analysis to future work.

Proof of Theorem

Step 1: Chain rule of KL divergence

$$\mathrm{KL}(p((A,B)\in\cdot)\|q((A,B)\in\cdot)) = \mathrm{KL}(p(A)\|q(A))\mathrm{KL}(p(B\in\cdot\mid A)\|q(B\in\cdot\mid A))$$

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+1:T}\in\cdot\mid\mathcal{D}_t,X_{t:T-1}\right)\|\mathbb{P}\left(Y_{t+1:T}\in\cdot\mid\theta_t,X_{t:T-1}\right)\right)\right]\\
=&\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+1}\in\cdot\mid\mathcal{D}_t,X_{t:T-1}\right)\|\mathbb{P}\left(Y_{t+1}\in\cdot\mid\theta_t,X_{t:T-1}\right)\right)\right]\\
&+\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+2:T}\in\cdot\mid\mathcal{D}_t,X_t,Y_{t+1},X_{t+1:T-1}\right)\|\mathbb{P}\left(Y_{t+2:T}\in\cdot\mid\theta_t,X_t,Y_{t+1},X_{t+1:T-1}\right)\right)\right]\\
=&\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+1}\in\cdot\mid\mathcal{D}_t,X_t\right)\|\mathbb{P}\left(Y_{t+1}\in\cdot\mid\theta_t,X_t\right)\right)\right]\\
&+\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+2:T}\in\cdot\mid\mathcal{D}_{t+1},X_{t+1:T-1}\right)\|\mathbb{P}\left(Y_{t+2:T}\in\cdot\mid\theta_t,X_t,Y_{t+1},X_{t+1:T-1}\right)\right)\right]
\end{aligned}$$

Proof of Theroem

Step 2: by lemma

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+2:T} \in \cdot \mid \mathcal{D}_{t+1}, X_{t+1:T-1}\right) \| \mathbb{P}\left(Y_{t+2:T} \in \cdot \mid \theta_t, X_t, Y_{t+1}, X_{t+1:T-1}\right)\right)\right] \\
&\leqslant \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+2:T} \in \cdot \mid \mathcal{D}_{t+1}, X_{t+1:T-1}\right) \| \mathbb{P}\left(Y_{t+2:T} \in \cdot \mid \theta_{t+1}, X_{t+1:T-1}\right)\right)\right]
\end{aligned}
$$

Proof of Theorem

Then, we have recursive computation,

$$
\begin{aligned}
&\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+1:T} \in \cdot \mid \mathcal{D}_t, X_{t:T-1}\right) \| \mathbb{P}\left(Y_{t+1:T} \in \cdot \mid \theta_t, X_{t:T-1}\right)\right)\right] \\
&\overset{(b)}{\leqslant} \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+1} \in \cdot \mid \mathcal{D}_t, X_t\right) \| \mathbb{P}\left(Y_{t+1} \in \cdot \mid \theta_t, X_t\right)\right)\right] \\
&+\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+2:T} \in \cdot \mid \mathcal{D}_{t+1}, X_{t+1:T-1}\right) \| \mathbb{P}\left(Y_{t+2:T} \in \cdot \mid \theta_{t+1}, X_{t+1:T-1}\right)\right)\right] \\
&\leqslant \ldots \\
&\leqslant \mathbb{E}\left[\sum_{t'=t}^{T-1} \mathbf{d}_{KL}\left(\mathbb{P}\left(Y_{t'+1} \in \cdot \mid \mathcal{D}_{t'}, X_{t'}\right) \| \mathbb{P}\left(Y_{t'+1} \in \cdot \mid \theta_{t'}, X_{t'}\right)\right)\right]
\end{aligned}
$$

Proof of Theorem

Step 3: by lemma

$$\begin{aligned}
&\mathbf{d}_{KL}\left(\mathbb{P}\left(Y_{t'+1} \in \cdot \mid \mathcal{D}_{t'}, X_{t'}\right) \| \mathbb{P}\left(Y_{t'+1} \in \cdot \mid \theta_{t'}, X_{t'}\right)\right) \\
&\leqslant \mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t'+1} \in \cdot \mid \mathcal{D}_{t'}, X_{t'}\right) \| \mathbb{P}\left(\hat{Y}_{t'+1} \in \cdot \mid \theta_{t'}, X_{t'}\right)\right)
\end{aligned}$$

Finally,

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t+1:T} \in \cdot \mid \mathcal{D}_t, X_{t:T-1}\right) \| \mathbb{P}\left(Y_{t+1:T} \in \cdot \mid \theta_t, X_{t:T-1}\right)\right)\right] \\
&\overset{(c)}{\leqslant} \mathbb{E}\left[\sum_{t'=t}^{T-1} \mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(Y_{t'+1} \in \cdot \mid \mathcal{D}_{t'}, X_{t'}\right) \| \mathbb{P}\left(\hat{Y}_{t'+1} \in \cdot \mid \theta_{t'}, X_{t'}\right)\right)\right] \\
&\overset{(d)}{=} \mathbb{E}\left[\sum_{t'=t}^{T-1} \mathbf{d}_{\mathrm{KL}}\left(\bar{P}_{t'+1} \| \hat{P}_{t'+1}\right)\right] \leqslant \epsilon,
\end{aligned}$$

Outline

Problem Setup

- Each time $t = 0, 1, \ldots$, the agent select action A_t and observes an outcome Y_t produced by the environments.
- Conditioned on environment $\mathcal{E}$ and action A_t, the next observation

$$Y_{t+1} \sim \mathcal{E}(\cdot \mid A_t)$$

- Real-valued reward function r encodes the preference of the agent over the observations,
- Objective:

$$\mathbb{E}\left[\sum_{t=0}^{T-1} r(Y_{t+1})\right]$$

Marginal predictions is not sufficient

- At any time step t, the rewards and observations at future time steps $t' > t$ are coupled through an unknown environment $\mathcal{E}$.
- Informative structure

Concrete example

- Bernoulli bandit with K independent action where

$$r(Y_{t+1}) = Y_{t+1}$$

- First $K-1$ actions, the agent knows the reward is distributed as $\mathrm{Bernoulli}(0.5)$
- While the final action produces the the deterministic outcome of 0 or 1, but it is equally likely to be of either kind.
- Best policy:
  - First select the final action to see if it is the optimal
  - and based on the first outcome, choose the arm that maximize expected reward given full knowledge of $\mathcal{E}$ for all future steps.

Relating joint predictions to regret

- To simplify exposition, we consider predictions for a vector of outcomes, $\mathbf{Y} \in \mathbb{R}^K$, with each entry corresponding to the outcome of an action.
- Relate the quality of future predictions about $\mathbf{Y}$ to agent performance on a Bernoulli bandit with correlated arms.
- K-armed Bernoulli bandit: $\mathcal{E} = \{p = (p_1, \ldots, p_K)\}$, where $p_k \in [0, 1]$ is the expected reward of k-th action.
- No assumptions on the prior $\mathbb{P}(p \in \cdot)$.
- Define the history by time t as $H_t = (A_0, Y_1, \ldots, A_{t-1}, Y_t)$.

Sequence of reward vectors from environment and from agent

- $\tilde{\mathbf{Y}}_{1:\tau}$ denote a sequence of τ vectors sampled from the environment $\mathcal{E}$. These τ vectors are conditionally independent given $\mathcal{E}$.
- Each vector has dimension K and the k-th component of each vector is conditionally independently sampled from $\mathrm{Bernoulli}(p_k)$. On the other hand, consider an agent that can also generate a sequence of K-dimensional binary vectors at each time t.
- Consider and agent that can also generate a sequence of K-dimensional binary vectors at each time t.
- Let θ_t denote its parameters and $\hat{\mathbf{Y}}^t_{1:\tau}$ denote a sequence of τ binary vectors sampled from it.

Sequence of reward vectors from environment and from agent

- $\tilde{\mathbf{Y}}_{1:\tau}$ denote a sequence of τ vectors sampled from the environment $\mathcal{E}$. These τ vectors are conditionally independent given $\mathcal{E}$.
- Each vector has dimension K and the k-th component of each vector is conditionally independently sampled from $\mathrm{Bernoulli}(p_k)$. On the other hand, consider an agent that can also generate a sequence of K-dimensional binary vectors at each time t.
- Consider and agent that can also generate a sequence of K-dimensional binary vectors at each time t.
- Let θ_t denote its parameters and $\hat{\mathbf{Y}}^t_{1:\tau}$ denote a sequence of τ binary vectors sampled from it.

Approximate Thompson sampling

Algorithm 1 Approximate Thompson sampling

Input: prior over environment parameters p
agent architecture
agent parameter initialization/update procedure
Initialization: compute parameters θ_0 based on prior
for $t = 0, 1, 2, \ldots$ **do**
  sample $\hat{\mathbf{Y}}^t_{1:\tau} \sim \mathbb{P}(\hat{\mathbf{Y}}_{1:\tau} \in \cdot | \theta_t)$
  sample $\hat{p}^t$ from $\mathbb{P}(p \in \cdot | \tilde{\mathbf{Y}}_{1:\tau} = \hat{\mathbf{Y}}^t_{1:\tau})$
  choose $A_t = \min \arg\max_k \hat{p}^t_k$
  compute θ_{t+1} based on θ_t and (A_t, Y_{t+1}).
end for

- $\min \arg\max_k \hat{p}^t_k$ is well defined. Specifically, $\arg\max_k \hat{p}^t_k \subseteq \{1, \ldots, K\}$ is a set.

Approximate Thompson sampling

- Note that Algorithm 1 is general in the sense that it does not depend on the agent's uncertainty representation.
- Instead, it only requires that the agent can simulate hypothetical observations, sampled from a joint predictive distribution.
- Also note that Algorithm 1 reduces to the standard (exact) Thompson sampling algorithm when $\mathbb{P}\left(\hat{\mathbf{Y}}_{1:\tau} \in \cdot \mid \theta_t\right) = \mathbb{P}\left(\tilde{\mathbf{Y}}_{1:\tau} \in \cdot \mid H_t\right)$ and $\tau \to \infty$.
- We use this algorithm to establish that an agent that performs well based on a particular loss function retains enough information to enable efficient exploration.

Regret bound

- (Bayes) cumulative regret: $\text{Regret}(T) = \sum_{t=0}^{T-1} \mathbb{E}\left[p_{A^*} - r\left(Y_{t+1}\right)\right]$, where $A^* = \min \arg\max_k p_k$ is one optimal action. Similarly, the expectation is over random outcomes, algorithmic randomness, and prior over $\mathcal{E}$.

Theorem 2.

For any integer $\tau \geqslant 1$ and any $\epsilon \in \Re_+$, if at each time t, the agent with parameters θ_t can generate samples $\hat{\mathbf{Y}}^t_{1:\tau}$ such that

$$\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\mathbb{P}\left(\tilde{\mathbf{Y}}_{1:\tau} \in \cdot \mid H_t\right) \| \mathbb{P}\left(\hat{\mathbf{Y}}^t_{1:\tau} \in \cdot \mid \theta_t\right)\right)\right] \leqslant \epsilon,$$

then under Algorithm 1, we have

$$\text{Regret}(T) \leqslant \sqrt{\frac{1}{2} K T \log K} + \left(\frac{K}{\sqrt{2\tau}} + \sqrt{2\epsilon}\right) T$$

Remark on the theorem

$$\text{Regret}(T) \leqslant \sqrt{\frac{1}{2}KT\log K} + \left(\frac{K}{\sqrt{2\tau}} + \sqrt{2\epsilon}\right)T$$

- First, note that if an agent can make good predictions $\tau \geqslant K/\epsilon$ steps into the future, then this regret bound reduces to $O(\sqrt{KT\log(K)} + \sqrt{K\epsilon}T)$, which is sufficient to ensure efficient exploration.
- Second, notice that this regret bound consists of three terms. The linear regret term $\sqrt{2\epsilon}T$ is due to the expected KL-divergence loss of the agent.
- Specifically, if the agent makes a perfect prediction in the sense that $\mathbb{P}\left(\hat{\mathbf{Y}}^t_{1:\tau} \in \cdot \mid \theta_t\right) = \mathbb{P}\left(\tilde{\mathbf{Y}}_{1:\tau} \in \cdot \mid H_t\right)$ for all t, then this linear regret term will reduce to zero.

Remark on the theorem

$$\mathrm{Regret}(T) \leqslant \sqrt{\frac{1}{2}KT\log K} + \left(\frac{K}{\sqrt{2\tau}} + \sqrt{2\epsilon}\right)T$$

- On the other hand, another linear regret term $KT/\sqrt{2\tau}$ is due to the fact that we choose $\tilde{A}$ as the learning target, which can $\tilde{A}$ be a sub-optimal action.
- It is obvious that as $\tau \to \infty, \tilde{A}$ will converge to A^* and this linear regret term will reduce to zero.
- Finally, the sublinear regret term $\sqrt{\frac{1}{2}KT\log K}$ is exactly the regret bound for the exact Thompson sampling algorithm (Russo & Van Roy, 2016).
- This is not surprising since when $\epsilon = 0$ (i.e. $\mathbb{P}\left(\hat{\mathbf{Y}}^t_{1:\tau} \in \cdot \mid \theta_t\right) = \mathbb{P}\left(\tilde{\mathbf{Y}}_{1:\tau} \in \cdot \mid H_t\right)$) and $\tau \to \infty$, Algorithm 1 reduces to the exact Thompson sampling algorithm.

Conjecture for more practical algorithm

- Note that in Algorithm 1, sampling $\hat{p}^t$ from $\mathbb{P}(p \in \cdot \mid \tilde{\mathbf{Y}}_{1:\tau} = \hat{\mathbf{Y}}^t_{1:\tau})$ can be computationally expensive.
- Instead, a computationally more efficient approach is to choose $\hat{p}^t$ as the sample mean of $\hat{\mathbf{Y}}_{1:\tau}$, i.e. $\hat{p}^t = \frac{1}{\tau}\sum_{i=1}^{\tau} \hat{\mathbf{Y}}^t_i$, where $\hat{\mathbf{Y}}^t_i$ is the i-th vector in $\hat{\mathbf{Y}}_{1:\tau}$.
- Conjecture: that one can derive a similar regret bound with this practical modification, but leave the analysis to future work.

Proof sketch of Theorem

- We provide a proof sketch for Theorem 5.1 in this subsection. First, notice that the expected per-step regret at time t is $\mathbb{E}\left[p_{A^*} - p_{A_t}\right]$, which can be decomposed as

$$\mathbb{E}\left[p_{A^*} - p_{A_t}\right] = \mathbb{E}\left[p_{A^*} - p_{\tilde{A}}\right] + \mathbb{E}\left[p_{\tilde{A}} - p_{A_t}\right]$$

- Recall that action $\tilde{A}$ is the learning target. We bound the two terms in the righthand side of equation (9) separately. First, based on the fact that p and $\tilde{p}$ are conditionally i.i.d given the environment proxy $\tilde{\mathbf{Y}}_{1:\tau}$, we can show that

$$\mathbb{E}\left[p_{A^*} - p_{\tilde{A}}\right] \leqslant K/\sqrt{2\tau}$$

Proof of the Theorem

- To bound the second term $\mathbb{E}\left[p_{\tilde{A}} - p_{A_t}\right]$, we consider its conditional version $\mathbb{E}_t\left[p_{\tilde{A}} - p_{A_t}\right]$, where the subscript t denotes conditioning on the history H_t. Using information-ratio analysis, we can prove that

$$\mathbb{E}_t\left[p_{\tilde{A}} - p_{A_t}\right] \leqslant \sqrt{\frac{K}{2}\mathbb{I}_t\left(\tilde{A}; A_t, \mathbf{Y}_{A_t}\right)} + \left\|\mathbb{P}_t(\tilde{A} \in \cdot) - \mathbb{P}_t\left(A_t \in \cdot\right)\right\|_1$$

- Using Pinsker's inequality, the data processing inequality, and the assumption on the expected KL-divergence in Theorem, we can bound that

$$\mathbb{E}\left[\left\|\mathbb{P}_t(\tilde{A} \in \cdot) - \mathbb{P}_t\left(A_t \in \cdot\right)\right\|_1\right] \leqslant \sqrt{2\epsilon}.$$

Proof of the Theorem

- On the other hand, based on Cauchy-Schwartz inequality and the chain rule for mutual information, we have

$$\sum_{t=0}^{T-1} \mathbb{E}\left[\sqrt{\mathbb{I}_t\left(\tilde{A}; A_t, \mathbf{Y}_{A_t}\right)}\right] \leqslant \sqrt{T\mathbb{I}\left(\tilde{A}; H_T\right)}$$

Finally, note that $\mathbb{I}\left(\tilde{A}; H_T\right) \leqslant \mathbb{H}(\tilde{A}) \leqslant \log K$. Combining the above inequalities, we have proved Theorem.

Outline

Outline

What is still missing?

-

References I

I. Osband, Z. Wen, S. M. Asghari, V. Dwaracherla, B. Hao, M. Ibrahimi, D. Lawson, X. Lu, B. O'Donoghue, and B. Van Roy. The Neural Testbed: Evaluating Joint Predictions. arXiv e-prints, art. arXiv:2110.04629, Oct. 2021.

Z. Wen, I. Osband, C. Qin, X. Lu, M. Ibrahimi, V. Dwaracherla, M. Asghari, and B. Van Roy. From predictions to decisions: The importance of joint predictive distributions. arXiv e-prints, pages arXiv–2107, 2021.