Efficient Gaussian Process Bandits via Believing only Informative Actions

Alec Koppel
U.S. Army Research Laboratory
Joint with: Amrit Singh Bedi (ARL), Dheeraj Peddireddy (Purdue), Vaneet Aggarwal (Purdue)

Uncertainty Representations in Bandits and Reinforcement Learning
INFORMS Annual Meeting
Nov. 10, 2020

Multi-Armed Bandits (MAB)

Setting: $\mathcal{A} = \{1, \ldots, A\}$ possible *arms* (actions) [1]

$\rightarrow$ suppose we select arm a_t at time t

$\Rightarrow$ then reward $r_t(a_t)$ is sequentially revealed by environment.

$\Rightarrow$ $r_t(a_t) \sim$ unknown distribution

$\rightarrow$ e.g., Bernoulli with unknown mean

[1] Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. SIAM journal on computing, 32(1), 48-77.
Bubeck, S., & Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. Machine Learning, 5(1), 1-122.

Multi-Armed Bandits (MAB)

Setting: $\mathcal{A} = \{1, \ldots, A\}$ possible *arms* (actions) [1]

$\rightarrow$ suppose we select arm a_t at time t

$\Rightarrow$ then reward $r_t(a_t)$ is sequentially revealed by environment.

$\Rightarrow$ $r_t(a_t) \sim$ unknown distribution

$\rightarrow$ e.g., Bernoulli with unknown mean

$\rightarrow$ **Goal**: select $\{a_t\}$ $\Rightarrow$ max. cumulative return $\sum_{t=1}^{\infty} r_t(a_t)$

$\Rightarrow$ want *regret* growth sublinearly in T, i.e., $\textbf{Reg}_T / T \rightarrow 0$ w/ T

$$\textbf{Reg}_T = \max_{\pi} \sum_{t=1}^{T} r_t(\pi) - \sum_{t=1}^{T} r_t(a_t) \quad \Rightarrow \text{in some probabilistic sense}$$

[1] Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. SIAM journal on computing, 32(1), 48-77.
Bubeck, S., & Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. Machine Learning, 5(1), 1-122.

Multi-Armed Bandits (MAB)

Setting: $\mathcal{A} = \{1, \ldots, A\}$ possible *arms* (actions) [1]

$\rightarrow$ suppose we select arm a_t at time t

$\Rightarrow$ then reward $r_t(a_t)$ is sequentially revealed by environment.

$\Rightarrow$ $r_t(a_t) \sim$ unknown distribution

$\rightarrow$ e.g., Bernoulli with unknown mean

$\rightarrow$ **Goal**: select $\{a_t\}$ $\Rightarrow$ max. cumulative return $\sum_{t=1}^{\infty} r_t(a_t)$

$\Rightarrow$ want *regret* growth sublinearly in T, i.e., $\mathbf{Reg}_T/T \rightarrow 0$ w/ T

$$\mathbf{Reg}_T = \max_{\pi} \sum_{t=1}^{T} r_t(\pi) - \sum_{t=1}^{T} r_t(a_t) \quad \Rightarrow \text{in some probabilistic sense}$$

$\pi = \pi_\theta \mapsto \mathcal{A}$ is time-invariant,.e.g, parameterized by $\theta \in \mathbb{R}^d$

[1] Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (2002). The nonstochastic multiarmed bandit problem. SIAM journal on computing, 32(1), 48-77.
Bubeck, S., & Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. Machine Learning, 5(1), 1-122.

Technological Context

Healthcare [2]:

$\rightarrow$ clinical trials: arm $\Rightarrow$ patient choice,reward $\Rightarrow$treatment efx.

$\rightarrow$ personalized dosing: arm $\Rightarrow$ dosage, reward $\Rightarrow$ no side efx.

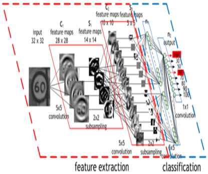
Recommender systems [3]:

$\rightarrow$ arm $\Rightarrow$ ad/article/movie on page, reward $\Rightarrow$ time on page

$\rightarrow$ dynamic pricing: arm $\Rightarrow$ price, reward $\Rightarrow$ revenue

Hyperparameter Search in Machine Learning Models [4]:

$\rightarrow$ arm $\Rightarrow$ regularizer/step-size, reward $\Rightarrow$ validation accuracy



[2] Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In Machine Learning for Healthcare Conference, pages 67–82, 2018

[3] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. CoRR, 2010.

[4] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. The Journal of Machine Learning Research, 18(1), 6765-6816.

Technological Context

Healthcare [2]:

→ clinical trials: arm ⇒ patient choice,reward ⇒treatment efx.

→ personalized dosing: arm ⇒ dosage, reward ⇒ no side efx.

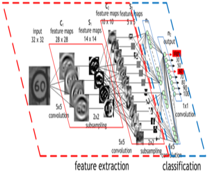
Recommender systems [3]:

→ arm ⇒ ad/article/movie on page, reward ⇒ time on page

→ dynamic pricing: arm ⇒ price, reward ⇒ revenue

Hyperparameter Search in Machine Learning Models [4]:

→ arm ⇒ regularizer/step-size, reward ⇒ validation accuracy



[2] Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In Machine Learning for Healthcare Conference, pages 67–82, 2018

[3] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. CoRR, 2010.

[4] Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. The Journal of Machine Learning Research, 18(1), 6765-6816.

Conceptual Context

Classical solutions to MAB $\Rightarrow$ track statistics of $r_t(a)$ for each a

$\rightarrow$ UCB [5]: track mean & std. dev., actions

$\Rightarrow$ select arms according to upper conf. bound

$$a_t = \operatorname*{argmax}_{a \in \mathcal{A}} \hat{\mu}_t + \beta \hat{\sigma}_t$$

[5] Lai, T. L. (1987). "Adaptive treatment allocation and the multi-armed bandit problem". The Annals of Statistics, 1091-1114. Lai, T. L., & Robbins, H. (1985). "Asymptotically efficient adaptive allocation rules." Advances in applied mathematics, 6(1), 4-22.

[6] Wang, Z., & de Freitas, N. Theoretical analysis of Bayesian optimisation with unknown Gaussian process hyper-parameters. NIPS Workshop on Bayesian Optimization, 2014

[7] Thompson, W. 1933. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". Biometrika. 25(3/4): 285–294.

[8] Gittins, J. & D. Jones. 1979. "A dynamic allocation index for the discounted multiarmed bandit problem". Biometrika. 66(3): 561–565.

Conceptual Context

Classical solutions to MAB $\Rightarrow$ track statistics of $r_t(a)$ for each a

$\rightarrow$ UCB [5]: track mean & std. dev., actions

$\Rightarrow$ select arms according to upper conf. bound

$$a_t = \operatorname*{argmax}_{a \in \mathcal{A}} \hat{\mu}_t + \beta \hat{\sigma}_t$$

$\rightarrow$ Expected Improvement [6] operates similarly

$\rightarrow$ Thompson Sampling [7] & Gittins index [8]:

$\Rightarrow$ construct distribution over rewards

$\Rightarrow$ select max estimated cond. mean of cumulative return

---

[5] Lai, T. L. (1987). "Adaptive treatment allocation and the multi-armed bandit problem". The Annals of Statistics, 1091-1114. Lai, T. L., & Robbins, H. (1985). "Asymptotically efficient adaptive allocation rules." Advances in applied mathematics, 6(1), 4-22.

[6] Wang, Z., & de Freitas, N. Theoretical analysis of Bayesian optimisation with unknown Gaussian process hyper-parameters. NIPS Workshop on Bayesian Optimization, 2014

[7] Thompson, W. 1933. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". Biometrika. 25(3/4): 285–294.

[8] Gittins, J. & D. Jones. 1979. "A dynamic allocation index for the discounted multiarmed bandit problem". Biometrika. 66(3): 561–565.

Conceptual Context

Classical solutions to MAB $\Rightarrow$ track statistics of $r_t(a)$ for each a

$\rightarrow$ UCB [5]: track mean & std. dev., actions

$\Rightarrow$ select arms according to upper conf. bound

$$a_t = \underset{a \in \mathcal{A}}{\text{argmax}}\, \hat{\mu}_t + \beta \hat{\sigma}_t$$

$\rightarrow$ Expected Improvement [6] operates similarly

$\rightarrow$ Thompson Sampling [7] & Gittins index [8]:

$\Rightarrow$ construct distribution over rewards

$\Rightarrow$ select max estimated cond. mean of cumulative return

$\rightarrow$ These approaches yield sublinear regret $\mathbf{Reg}_T / T \rightarrow 0$

$\Rightarrow$ exhibit computational challenges when # arms A large

$\Rightarrow$ that is, need statistics/posterior with complexity $\propto A$ or T

---

[5] Lai, T. L. (1987). "Adaptive treatment allocation and the multi-armed bandit problem". The Annals of Statistics, 1091-1114. Lai, T. L., & Robbins, H. (1985). "Asymptotically efficient adaptive allocation rules." Advances in applied mathematics, 6(1), 4-22.

[6] Wang, Z., & de Freitas, N. Theoretical analysis of Bayesian optimisation with unknown Gaussian process hyper-parameters. NIPS Workshop on Bayesian Optimization, 2014

[7] Thompson, W. 1933. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". Biometrika. 25(3/4): 285–294.

[8] Gittins, J. & D. Jones. 1979. "A dynamic allocation index for the discounted multiarmed bandit problem". Biometrika. 66(3): 561–565.

Time and Action Space Complexity

When $\mathcal{A}$ is either continuous, or discrete but A is large scale

$\rightarrow$ tracking conditional mean/variance or density is costly

$\rightarrow$ **Lipschitz bandits**: discretize space & form bins

$\Rightarrow$ balance regret, number of parameters = # bins $\propto$ T [9] [10]

[9] Magureanu, S., Combes, R., & Proutiere, A. (2014, May). Lipschitz Bandits: Regret Lower Bound and Optimal Algorithms. In Conference on Learning Theory (pp. 975-999).

[10] Bubeck, S., Stoltz, G., & Yu, J. Y. (2011, October). Lipschitz bandits without the Lipschitz constant. In International Conference on Algorithmic Learning Theory (pp. 144-158). Springer, Berlin, Heidelberg.

[11] Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. W. (2012). Information-theoretic regret bounds for gaussian process optimization in the bandit setting. IEEE Transactions on Information Theory, 58(5), 3250-3265.

[12] Wang, Z., & de Freitas, N. Theoretical analysis of Bayesian optimisation with unknown Gaussian process hyper-parameters. NIPS Workshop on Bayesian Optimization, 2014

[13] Nguyen, V., Gupta, S., Rana, S., Li, C., & Venkatesh, S. (2017, November). Regret for expected improvement over the best-observed value and stopping condition. In Asian Conference on Machine Learning (pp. 279-294).

Time and Action Space Complexity

When $\mathcal{A}$ is either continuous, or discrete but A is large scale

$\rightarrow$ tracking conditional mean/variance or density is costly

$\rightarrow$ **Lipschitz bandits**: discretize space & form bins

$\Rightarrow$ balance regret, number of parameters = # bins $\propto$ T [9] [10]

$\rightarrow$ **Gaussian Process** $\Rightarrow$ define posterior directly over $\mathcal{A}$

$\Rightarrow$ posterior dist. used to define UCB [11], EI[12], MPI[13]

$\Rightarrow$ complexity of computing posterior parameters $\mathcal{O}(T^3)$

$\rightarrow$ **Central question**: can we define no-regret bandit algorithm

$\Rightarrow$ whose complexity remains moderate as $A, T \rightarrow \infty$

---

9 Magureanu, S., Combes, R., & Proutiere, A. (2014, May). Lipschitz Bandits: Regret Lower Bound and Optimal Algorithms. In Conference on Learning Theory (pp. 975-999).

10 Bubeck, S., Stoltz, G., & Yu, J. Y. (2011, October). Lipschitz bandits without the Lipschitz constant. In International Conference on Algorithmic Learning Theory (pp. 144-158). Springer, Berlin, Heidelberg.

11 Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. W. (2012). Information-theoretic regret bounds for gaussian process optimization in the bandit setting. IEEE Transactions on Information Theory, 58(5), 3250-3265.

12 Wang, Z., & de Freitas, N. Theoretical analysis of Bayesian optimisation with unknown Gaussian process hyper-parameters. NIPS Workshop on Bayesian Optimization, 2014

13 Nguyen, V., Gupta, S., Rana, S., Li, C., & Venkatesh, S. (2017, November). Regret for expected improvement over the best-observed value and stopping condition. In Asian Conference on Machine Learning (pp. 279-294).

Gaussian Processes

GPs $\Rightarrow$ nonparametric Bayesian method ($\mathcal{A} \subset \mathbb{R}^p$, $\mathcal{Y} \subset \mathbb{R}$)

$\Rightarrow \hat{y} = f(\mathbf{a})$ $\Rightarrow$ capture relationship of $(\mathbf{a}, y) \in \mathcal{A} \times \mathcal{Y}$

$\Rightarrow$ estimate f via $T - 1$ training examples $\mathcal{S} = \{\mathbf{a}_t, y_t\}_{t=1}^{T-1}$.

$\rightarrow$ Suppose $f(\mathbf{a})$ in a parameterized family $\Rightarrow$ estimate params

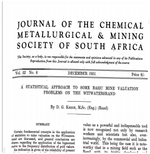
$\rightarrow$ *Prior* $\mathbf{f}_{\mathcal{S}} = [f(\mathbf{a}_t), \cdots, f(\mathbf{a}_{t-1})]$ $\Rightarrow$ Gaussian: $\mathbf{f}_{\mathcal{S}} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{t-1})$

$\Rightarrow$ Covariance $\mathbf{K}_{T-1} = [\kappa(\mathbf{a}_{t'}, \mathbf{a}_t)]_{m,t=1}^{T-1,T-1}$ via kernel $\kappa : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}$

$\Rightarrow$ Kernel $\Rightarrow$ distance, e.g., $\kappa(\mathbf{a}_{t'}, \mathbf{a}_t) = \exp\{-\|\mathbf{a}_{t'} - \mathbf{a}_t\|^2 / c^2\}$

$\rightarrow$ Standard GP $\Rightarrow$ $\mathbf{f}_{\mathcal{S}}$ observed in noise $\mathbb{P}(\mathbf{y} \mid \mathbf{f}_{\mathcal{S}}) = \mathcal{N}(\mathbf{f}_{\mathcal{S}}, \sigma^2 \mathbf{I})$

$\Rightarrow$ where σ^2 is some variance parameter.

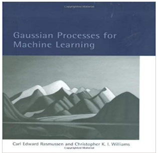

JOURNAL OF THE CHEMICAL METALLURGICAL & MINING SOCIETY OF SOUTH AFRICA

Vol. 52 No. 6 DECEMBER 1951 Price 6/-

A STATISTICAL APPROACH TO SOME BASIC MINE VALUATION PROBLEMS ON THE WITWATERSRAND

By D. G. Krige, M.Sc. (Eng.) (Rand)

SUMMARY



Gaussian Processes

GPs $\Rightarrow$ nonparametric Bayesian method ($\mathcal{A} \subset \mathbb{R}^p$, $\mathcal{Y} \subset \mathbb{R}$)

$\Rightarrow$ $\hat{y} = f(\mathbf{a})$ $\Rightarrow$ capture relationship of $(\mathbf{a}, y) \in \mathcal{A} \times \mathcal{Y}$

$\Rightarrow$ estimate f via $T-1$ training examples $\mathcal{S} = \{\mathbf{a}_t, y_t\}_{t=1}^{T-1}$.

$\rightarrow$ Suppose $f(\mathbf{a})$ in a parameterized family $\Rightarrow$ estimate params

$\rightarrow$ Upon receiving new sample $\mathbf{a}_t$, form posterior for $\hat{y}_t$ as

$$\mathbb{P}(y_t \,|\, \mathcal{S} \cup \mathbf{a}_t) = \mathcal{N}\left(\boldsymbol{\mu}_{t \,|\, \mathcal{S}}, \boldsymbol{\Sigma}_{t \,|\, \mathcal{S}}\right)$$

$\Rightarrow$ where the mean and covariance are given by

$$\begin{aligned}\boldsymbol{\mu}_{t \,|\, \mathcal{S}} &= \mathbf{k}_{\mathcal{S}}(\mathbf{a}_t)[\mathbf{K}_{T-1} + \sigma^2 \mathbf{I}]^{-1} \mathbf{y}_{T-1} \\ \boldsymbol{\Sigma}_{t \,|\, \mathcal{S}} &= \kappa(\mathbf{a}_t, \mathbf{a}_t) \\ &\qquad - \mathbf{k}_{\mathcal{S}}^T(\mathbf{a}_t)[\mathbf{K}_{T-1} + \sigma^2 \mathbf{I}]^{-1} \mathbf{k}_{\mathcal{S}}(\mathbf{a}_t)\end{aligned}$$

$\Rightarrow$ $\mathbf{k}_{\mathcal{S}}(\mathbf{a}) = [\kappa(\mathbf{a}_1, \mathbf{a}); \cdots \kappa(\mathbf{a}_{t-1}, \mathbf{a})]$ $\Rightarrow$ empirical kernel map

Gaussian Process Bandits

Gaussian Process posterior $\Rightarrow$ use w/ various action selections

$\rightarrow$ incorporates all past samples into posterior at present

$\Rightarrow$ Upper-Confidence Bound (UCB)[11]: $\Rightarrow$ via posterior stats.

$$\mathbf{a}_{t+1} = \underset{\mathbf{a}\in\mathcal{A}}{\text{argmax}} \underbrace{\boldsymbol{\mu}_{T \,|\, \mathcal{S}_{T-1}} + \sqrt{\beta_t \boldsymbol{\Sigma}_{T \,|\, \mathcal{S}_{T-1}}}}_{:=\alpha_{\text{UCB}}(\mathbf{a})}$$

$\rightarrow$ Expected Improvement (EI)[13]: $\Rightarrow$ action selected as

$$\mathbf{a}_{t+1} = \underset{\mathbf{a}\in\mathcal{A}}{\text{argmax}} \underbrace{\sigma_{t-1}(\mathbf{a})\phi(z) + [\mu_{t-1}(\mathbf{a}) - y_{t-1}^{\max}]\Phi(z)}_{:=\alpha_{\text{EI}}(\mathbf{a})} ,$$

$\Rightarrow$ where $y_{t-1}^{\max} = \max\{y_u\}_{u\leq t}$

$\Rightarrow$ $z = z_{t-1}(\mathbf{a}) = (\mu_{t-1}(\mathbf{a}) - y_{t-1}^{\max})/\sigma_{t-1}(\mathbf{a})$

$\Rightarrow$ $\phi(z)/\Phi(z)$ denote density/distribution standard Gaussian

Curse of Dimensionality

Posterior $\Rightarrow$ complexity scales at least $\mathcal{O}(T)$

$\Rightarrow$ As time $T \to \infty$ $\Rightarrow$ GP bandits require infinite complexity

$\to$ Approaches to compress GPs [14]

$\Rightarrow$ forward selection [15]

$\Rightarrow$ variational approx. GP likelihood [16]

$\to$ Fix memory M,"project" onto fixed "subspace," may **diverge**

---

[14] Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning (Vol. 2, No. 3, p. 4). MIT press.

[15] Csat, L., & Opper, M. (2002). Sparse on-line Gaussian processes. Neural computation, 14(3), 641-668.
Seeger, Matthias, Christopher Seeger, Matthias, Christopher KI Williams, and Neil D. Lawrence. "Fast forward selection to speed up sparse gaussian process regression." Workshop on AI and Stats. 9. 2003.

[16] Titsias, M. (2009, April). Variational learning of inducing variables in sparse Gaussian processes. In AISTATS (pp. 567-574).
Snelson, E., & Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In NeurIPS (pp. 1257-1264).

[17] Calandriello, D., Carratino, L., Lazaric, A., Valko, M., & Rosasco, L. (2019, June). Gaussian Process Optimization with Adaptive Sketching: Scalable and No Regret. In Conference on Learning Theory (pp. 533-557).

[18] Chowdhury, S. R., & Gopalan, A. (2019, April). Online learning in kernelized markov decision processes. AISTATS (pp. 3197-3205).

[19] Calandriello, D., Carratino, L., Valko, M., Lazaric, A., & Rosasco, L. (2020). Near-linear Time Gaussian Process Optimization with Adaptive Batching and Resparsification. arXiv preprint arXiv:2002.09954.

Curse of Dimensionality

Posterior $\Rightarrow$ complexity scales at least $\mathcal{O}(T)$

$\Rightarrow$ As time $T \to \infty$ $\Rightarrow$ GP bandits require infinite complexity

$\to$ Approaches to compress GPs [14]

$\Rightarrow$ forward selection [15]

$\Rightarrow$ variational approx. GP likelihood [16]

$\to$ Fix memory M,"project" onto fixed "subspace," may **diverge**

$\to$ **Key Idea**: posterior grows/shrinks w.r.t. data importance

$\Rightarrow$ quantified by cond. entropy $\Rightarrow$ information-theoretic regret

$\Rightarrow$ directly tunable tradeoff between model complexity/regret

$\to$ Alternatives use eff. prob. dim., an offline statistic [17] [18] [19]

---

14 Williams, C. K., & Rasmussen, C. E. (2006). Gaussian processes for machine learning (Vol. 2, No. 3, p. 4). MIT press.

15 Csat, L., & Opper, M. (2002). Sparse on-line Gaussian processes. Neural computation, 14(3), 641-668. Seeger, Matthias, Christopher Seeger, Matthias, Christopher KI Williams, and Neil D. Lawrence. "Fast forward selection to speed up sparse gaussian process regression." Workshop on AI and Stats. 9. 2003.

16 Titsias, M. (2009, April). Variational learning of inducing variables in sparse Gaussian processes. In AISTATS (pp. 567-574). Snelson, E., & Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. In NeurIPS (pp. 1257-1264).

17 Calandriello, D., Carratino, L., Lazaric, A., Valko, M., & Rosasco, L. (2019, June). Gaussian Process Optimization with Adaptive Sketching: Scalable and No Regret. In Conference on Learning Theory (pp. 533-557).

18 Chowdhury, S. R., & Gopalan, A. (2019, April). Online learning in kernelized markov decision processes. AISTATS (pp. 3197-3205).

19 Calandriello, D., Carratino, L., Valko, M., Lazaric, A., & Rosasco, L. (2020). Near-linear Time Gaussian Process Optimization with Adaptive Batching and Resparsification. arXiv preprint arXiv:2002.09954.

Proposed Algorithm

for t = 1,2... **do**

Select action $\mathbf{a}_t$ via UCB or EI :

$$\mathbf{a}_t = \arg\max_{\mathbf{a}\in\mathcal{X}} \alpha(\mathbf{a})$$

Sample: $y_t = f(\mathbf{a}_t) + \epsilon_t$, i.e., arm $r_t(a_t)$

If cond. entropy exceeds ϵ threshold $\mathbf{H}(y_t|\mathbf{y}_{t-1}) > \epsilon$

Augment dict. $\mathbf{D}_t = [\mathbf{D}_{t-1}; \mathbf{a}_t]$, append target $\mathbf{y}_{\mathbf{D}_t} = [\mathbf{y}_{\mathbf{D}_{t-1}}; y_t]$

Update posterior mean $\mu_{\mathbf{D}_t}(\mathbf{a})$ & variance $\sigma_{\mathbf{D}_t}(\mathbf{a})$

$$\mu_{\mathbf{D}_t}(\mathbf{a}) = \boldsymbol{k}_{\mathbf{D}_t}(\mathbf{a})^T(\mathbf{K}_{\mathbf{D}_t} + \sigma^2\mathbf{I})^{-1}\mathbf{y}_{\mathbf{D}_t}$$

$$\sigma^2_{\mathbf{D}_t}(\mathbf{a}) = \kappa(\mathbf{a},\mathbf{a}') - \boldsymbol{k}_{\mathbf{D}_t}(\mathbf{a})^T(\mathbf{K}_{\mathbf{D}_t,\mathbf{D}_t} + \sigma^2\mathbf{I})^{-1}\boldsymbol{k}_{\mathbf{D}_t}(\mathbf{a}')$$

else

Fix dict. $\mathbf{D}_t = \mathbf{D}_{t-1}$, target $\mathbf{y}_{\mathbf{D}_t} = \mathbf{y}_{\mathbf{D}_{t-1}}$, & GP.

$$(\mu_{\mathbf{D}_t}(\mathbf{a}), \sigma_{\mathbf{D}_t}(\mathbf{a}), \mathbf{D}_t) = (\mu_{\mathbf{D}_{t-1}}(\mathbf{a}), \sigma_{\mathbf{D}_{t-1}}(\mathbf{a}), \mathbf{D}_{t-1})$$

end for

Conditional entropy of GP can be evaluated in closed form as $\mathbf{H}(y_t|\mathbf{y}_{t-1}) = \frac{1}{2}\log\left(2\pi e(\sigma^2 + \sigma^2_{\mathbf{D}_{t-1}}(\mathbf{a}_t))\right)$.

Compression Intuition

Matrix of past actions $\mathbf{A}_t = [\mathbf{a}_1; \cdots; \mathbf{a}_{t-1}]$

$\Rightarrow$ dictionary $\mathbf{D}_t$ $\Rightarrow$ subset of columns

$\Rightarrow$ add $\mathbf{a}_t$ only if "significant"

If $\mathbf{H}(y_t|\hat{\mathbf{y}}_{t-1}) > \epsilon$

update $\mathbf{D}_t = [\mathbf{D}_{t-1}\ ;\ \mathbf{a}_t]$

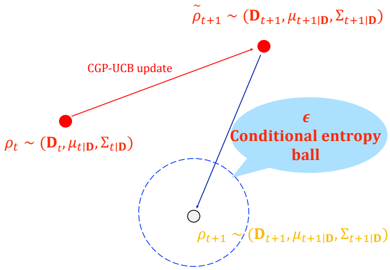
else

update $\mathbf{D}_t = \mathbf{D}_{t-1}$,

$\rightarrow$ where

$$\mathbf{H}(y_t|\hat{\mathbf{y}}_{t-1}) = \frac{1}{2}\log\left(2\pi e(\sigma^2 + \sigma^2_{\mathbf{D}_{t-1}}(\mathbf{a}_t))\right)$$

$\Rightarrow$ cond. entropy of GP



Regret Bound for GP-UCB

Theorem

For $\delta \in (0, 1)$, the proposed CGP-UCB achieves:

(i) for finite decision set

$$\mathbb{P}\left\{ \textbf{\textit{Reg}}_T \leq \sqrt{C_1 T \beta_T \hat{\gamma}_T} + \sqrt{\epsilon} T \right\} \geq 1 - \delta,$$

(ii) for general decision set

$$\mathbb{P}\left\{ \textbf{\textit{Reg}}_T \leq \sqrt{C_1 T \beta_T \hat{\gamma}_T} + \sqrt{\epsilon} T + \frac{\pi^2}{6} \right\} \geq 1 - \delta.$$

Theorem

Suppose that the entropy $H(\{y_t\})$ is bounded for all t. Then, the number of elements in the dictionary $\mathbf{D}_T$ denoted by $M_T(\epsilon)$ is finite as $T \to \infty$ for fixed compression threshold ϵ.

Regret Bound for GP-EI

Theorem

For the finite decision set, with $\delta \in (0,1)$, the proposed Compressed EI achieves

$$\mathbb{P}\left\{ \boldsymbol{Reg}_T \leq \sqrt{\frac{2T\left(\gamma_T + \epsilon T\right)}{\log\left(1+\sigma^{-2}\right)}} \left[\sqrt{3(\beta_T + 1 + R^2)} + \sqrt{\beta_T}\right] \right\} \geq 1 - \delta \,,$$

where

$$R := \sup_{t \geq 0} \sup_{\mathbf{a} \in \mathcal{X}} \frac{|\mu_{t-1}(\mathbf{a}) - y^{\max}|}{\sigma_{t-1}(\mathbf{a})}$$

Experimentation

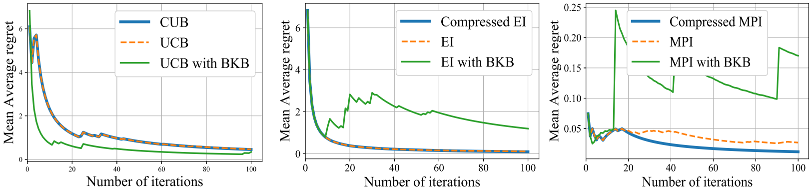
We employ proposed scheme for ML hyperparameter tuning

$\Rightarrow$ multi-class classification using CNN on MNIST data set

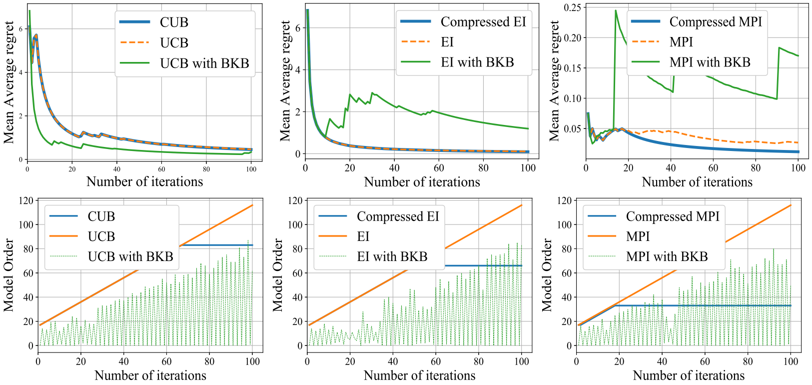
$\Rightarrow$ learning rate, batch size, dropout of inputs, ℓ_2 regularizer

$\rightarrow$ Instantaneous reward is statistical accuracy

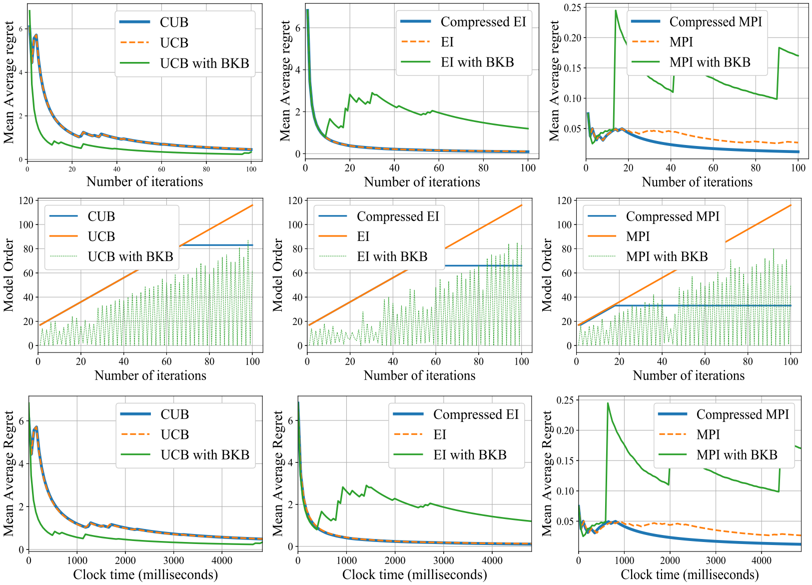
Hyper-parameter Tuning



Hyper-parameter Tuning



Hyper-parameter Tuning



Conclusion

We consider the problems with large actions spaces

$\rightarrow$ Parameterize the action distribution as the GP

$\rightarrow$ Unfortunately, GP exhibit complexity challenges

$\Rightarrow$ grows cubically with the time

$\Rightarrow$ Memory requirement grows indefinitely

$\rightarrow$ Designed entropy-base compression

$\Rightarrow$ balance regret and posterior complexity on the fly

$\rightarrow$ Future directions :

$\Rightarrow$ extensions to non-stationary (restless) bandits,

$\Rightarrow$ information theoretic compression of CNNs

References

⇒ A. S. Bedi, D. Peddireddy, V. Aggarwal, and A. Koppel, ”Efficient Large-Scale Gaussian Process Bandits by Believing only Informative Actions ,” in Learning for Dynamics and Control (L4DC), University of California, Berkeley, CA , June 2020.

→ A. S. Bedi, D. Peddireddy, V. Aggarwal, and A. Koppel, ”Efficient Gaussian Process Bandits by Believing only Informative Actions,” arXiv preprint (2020).