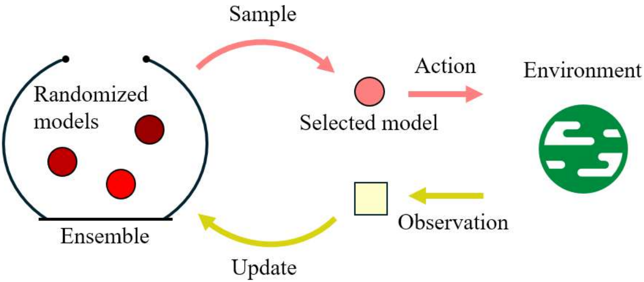
Improved Regret of Linear Ensemble Sampling

Harin Lee, Min-hwan Oh

Seoul National University

Ensemble Sampling

- Practically efficient randomized exploration strategy
- Maintains an ensemble of models, sample one to take an action.
- Has shown a significant improvement of performance in various tasks, including DQN (Osband et al., 2016, 2018, 2019).



Linear Ensemble Sampling

Linear Ensemble Sampling (Lu and Van Roy, 2017)

- Ensemble sampling for linear bandits
  - Linear bandit : Arm set $\mathcal{X} \subset \mathbb{R}^d$, reward $Y_t = X_t^\top \theta^* + \eta_t$

Algorithm Linear Ensemble Sampling

Initialize $\theta_0^1, \ldots, \theta_0^m \in \mathbb{R}^d$
for $t = 1, 2, \ldots T$ **do**
  Sample $j_t \in [m]$
  Pull arm $X_t = \mathrm{argmax}_{x \in \mathcal{X}}\, x^\top \theta_{t-1}^{j_t}$ and observe Y_t
  Update $\theta_t^1, \ldots, \theta_t^m \in \mathbb{R}^d$
end for

Linear Ensemble Sampling

Linear Ensemble Sampling (Lu and Van Roy, 2017)

- Ensemble sampling for linear bandits
  - Linear bandit : Arm set $\mathcal{X} \subset \mathbb{R}^d$, reward $Y_t = X_t^\top \theta^* + \eta_t$

Algorithm Linear Ensemble Sampling

Initialize $\theta_0^1, \ldots, \theta_0^m \in \mathbb{R}^d$
for $t = 1, 2, \ldots T$ **do**
  Sample $j_t \in [m]$
  Pull arm $X_t = \mathrm{argmax}_{x \in \mathcal{X}} x^\top \theta_{t-1}^{j_t}$ and observe Y_t
  Update $\theta_t^1, \ldots, \theta_t^m \in \mathbb{R}^d$
end for

- Sample perturbations $W^j \in \mathbb{R}^d$ and $\{Z_i^j\}_{i=1}^T$ for each $j \in [m]$.
- Each θ_t^j is the solution of the following minimization problem:

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}}\, \lambda \left\| \theta - W^j / \lambda \right\|_2^2 + \sum_{i=1}^{t} \left(X_i^\top \theta - \left(Y_i + Z_i^j \right) \right)^2$$

Linear Ensemble Sampling

Linear Ensemble Sampling (Lu and Van Roy, 2017)

- Ensemble sampling for linear bandits
  - Linear bandit : Arm set $\mathcal{X} \subset \mathbb{R}^d$, reward $Y_t = X_t^\top \theta^* + \eta_t$

Algorithm Linear Ensemble Sampling

Initialize $\theta_0^1, \ldots, \theta_0^m \in \mathbb{R}^d$
for $t = 1, 2, \ldots T$ **do**
  Sample $j_t \in [m]$
  Pull arm $X_t = \text{argmax}_{x \in \mathcal{X}} x^\top \theta_{t-1}^{j_t}$ and observe Y_t
  Update $\theta_t^1, \ldots, \theta_t^m \in \mathbb{R}^d$
end for

- Sample perturbations $W^j \in \mathbb{R}^d$ and $\{Z_i^j\}_{i=1}^T$ for each $j \in [m]$.
- Each θ_t^j is the solution of the following minimization problem:

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}}\, \lambda \left\| \theta - W^j / \lambda \right\|_2^2 + \sum_{i=1}^{t} \left(X_i^\top \theta - \left(Y_i + Z_i^j \right) \right)^2$$

- Incremental updates are possible for linear models and gradient descent-based models (e.g. neural net).

Existing Analysis of Ensemble Sampling

Table: Comparison of regret bounds for linear ensemble sampling.

Paper	Freq / Bayes	Regret Bound	Ensemble Size
Lu and Van Roy (2017)	Frequentist	Invalid	Invalid
Qin et al. (2022)	Bayesian	$\widetilde{\mathcal{O}}(\sqrt{dT \log K})$	$\Omega(KT)$
Janz et al. (2023)	Frequentist	$\widetilde{\mathcal{O}}(d^{5/2}\sqrt{T})$	$\Theta(d \log T)$
This work	Frequentist	$\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$	$\Omega(K \log T)$

The previous analyses were unsatisfactory since:

- Bayesian regret is weaker than frequentist regret.
- Ensemble size that scales linearly with T is impractical.
- $\widetilde{\mathcal{O}}(d^{5/2}\sqrt{T})$ is worse than $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ of Thompson sampling.

Main Result

Theorem : Regret bound of linear ensemble sampling)

Let $K < \infty$ be the number of arms. With $\lambda \geq 1$, $W^j \sim \mathcal{N}(\mathbf{0}_d, \lambda\beta_T^2 I_d)$, $Z_i^j \sim \mathcal{N}(0, \beta_T^2)$ and $m \geq \Omega(K \log T + \log \frac{1}{\delta})$, the regret of linear ensemble sampling is

$$\text{Regret}_T = \mathcal{O}((d \log T)^{3/2}\sqrt{T}) \, .$$

- $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ frequentist regret bound.
  - Improves the bound of Janz et al. (2023) by a factor of d.
  - Matches the best known regret bound of randomized algorithm (Abeille and Lazaric, 2017), up to log factors.
- Ensemble size logarithmic in T.

Regret Analysis (1/2)

Theorem : General regret bound for linear bandit algorithm

Assume that the agent chooses $X_t = \mathrm{argmax}_{x \in \mathcal{X}} x^\top \theta_t$ for some estimator θ_t. Let $V_t = \lambda I + \sum_{i=1}^t X_i X_i^\top$.

1. (Concentration) There exists a constant $\gamma > 0$ such that $\|\theta_t - \theta^*\|_{V_{t-1}} \leq \gamma$.
2. (Optimism) There exists a constant $p \in (0, 1]$ such that $\mathbb{P}\left(\left(x^{*\top}\theta^* \leq X_t^\top \theta_t\right) \mid \mathcal{F}_{t-1}\right) \geq p$.

Then, the cumulative regret of T time steps is

$$R(T) = \widetilde{\mathcal{O}}\left(\frac{\gamma}{p}\sqrt{dT}\right) .$$

- *Concentration* and *optimism* implies $O(\sqrt{T})$ regret.
- Simpler and more rigorous proof utilizing Markov's inequality

Regret Analysis (2/2)

Claim : Optimism for linear ensemble sampling

Let m be the size of ensemble and $p = 0.15$. There exists an event $\mathcal{E}^*$ under which at least mp models in the ensemble are optimistic, and $\mathbb{P}(\mathcal{E}^{*\mathsf{C}}) \leq T^K \exp(-m/C)$.

- Apply Hoeffding's inequality and the union bound over possible sequences of arms.
- Take ensemble size $m \geq C(K \log T + \log \frac{1}{\delta})$ so that the probability of failure is at most δ.

Relationship with Perturbed History Exploration

Linear Perturbed History Exploration (LinPHE) (Kveton et al., 2020)

- Samples fresh perturbation and recomputes perturbed estimator at each time step.

Proposition

Linear ensemble sampling with T models and round robin model selection rule is equivalent to LinPHE.

Corollary : Regret bound for LinPHE

LinPHE achieves $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret bound.

Summary

- We prove a $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret bound for linear ensemble sampling with ensemble of $\Omega(K \log T)$ models.
- The regret bound improves the previous result by a factor of d and matches the best bound of randomized algorithms.
- We introduce a novel analysis framework that holds for a wide variety of algorithms, which may be of independent interest.
- We rigorously demonstrate the relationship between linear ensemble sampling and LinPHE, which leads to a $\widetilde{\mathcal{O}}(d^{3/2}\sqrt{T})$ regret bound for LinPHE.

References I

Abeille, M. and Lazaric, A. (2017). Linear Thompson Sampling Revisited. In Singh, A. and Zhu, J., editors, Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of Proceedings of Machine Learning Research, pages 176–184. PMLR, PMLR.

Janz, D., Litvak, A. E., and Szepesvári, C. (2023). Ensemble sampling for linear bandits: small ensembles suffice. arXiv preprint arXiv:2311.08376.

Kveton, B., Szepesvári, C., Ghavamzadeh, M., and Boutilier, C. (2020). Perturbed-history exploration in stochastic linear bandits. In Uncertainty in Artificial Intelligence, pages 530–540. PMLR.

Lu, X. and Van Roy, B. (2017). Ensemble sampling. Advances in Neural Information Processing Systems, 30.

Osband, I., Aslanides, J., and Cassirer, A. (2018). Randomized prior functions for deep reinforcement learning. Advances in Neural Information Processing Systems, 31.

Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016). Deep exploration via bootstrapped dqn. Advances in Neural Information Processing Systems, 29.

Osband, I., Van Roy, B., Russo, D. J., Wen, Z., et al. (2019). Deep exploration via randomized value functions. Journal of Machine Learning Research, 20(124):1–62.

Qin, C., Wen, Z., Lu, X., and Van Roy, B. (2022). An analysis of ensemble sampling. Advances in Neural Information Processing Systems, 35:21602–21614.