A Survey of Risk-Aware Multi-Armed Bandits

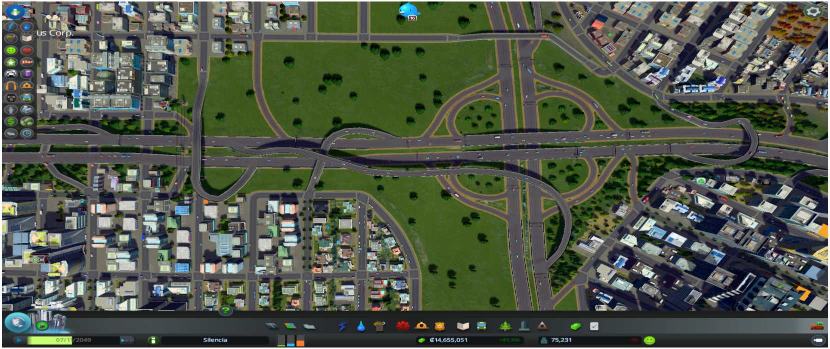
Vincent Tan | **Prashanth L. A.** | **Krishna Jagannathan**

NUS | IIT Madras | IIT Madras



IJCAI (Vienna, Austria), Jul 2022

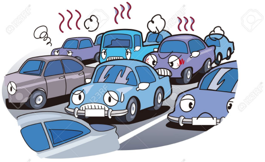
Going to office: Bandit style



On every day

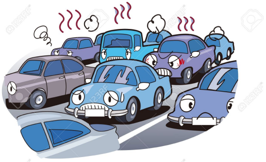
1. Pick a route to office
2. Reach office and record (suffered) delay

Why Consider Risk?



$\mathbb{E}[\text{time}] = 10 \text{ mins}, \Pr(\text{jam}) = 0.1$ $\mathbb{E}[\text{time}] = 11 \text{ mins}, \Pr(\text{jam}) = 0$

Why Consider Risk?



$\mathbb{E}[\text{time}] = 10 \text{ mins}, \Pr(\text{jam}) = 0.1$ $\quad$ $\mathbb{E}[\text{time}] = 11 \text{ mins}, \Pr(\text{jam}) = 0$

- Delays are stochastic.
- In choosing between routes, we need not necessarily minimize expected delay.
- Two route scenario: Average delay of Route 1 slightly below that of Route 2.
- Route 1 has a small chance of very high delay, e.g., jams.
- I might prefer Route 2.

Preliminary Definitions I

Definition

Given i.i.d. random samples $\{X_i\}_{i=1}^n$ from the distribution of a random variable, the empirical distribution function is

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\} \quad \text{for any} \quad x \in \mathbb{R}.$$

Preliminary Definitions I

Definition

Given i.i.d. random samples $\{X_i\}_{i=1}^n$ from the distribution of a random variable, the empirical distribution function is

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \le x\} \quad \text{for any} \quad x \in \mathbb{R}.$$

Definition

Random variable X is σ^2-sub-Gaussian if its cumulant generating function

$$\log \mathbb{E}[\exp(rX)] \le \frac{r^2\sigma^2}{2} \quad \text{for all} \quad r \in \mathbb{R}.$$

See Wainwright (2019, Theorem 2.1) for equivalent characterizations.

Preliminary Definitions I

Definition

The Wasserstein distance between two cumulative distribution functions (CDFs) F_1 and F_2 on $\mathbb{R}$ is

$$W_1(F_1, F_2) := \inf_{F \in \Gamma(F_1, F_2)} \int_{\mathbb{R}^2} |x - y| \, \mathrm{d}F(x, y)$$

where $\Gamma(F_1, F_2)$ is the set of couplings of F_1 and F_2.

Preliminary Definitions I

Definition

The Wasserstein distance between two cumulative distribution functions (CDFs) F_1 and F_2 on $\mathbb{R}$ is

$$W_1(F_1, F_2) := \inf_{F \in \Gamma(F_1, F_2)} \int_{\mathbb{R}^2} |x - y| \, \mathrm{d}F(x, y)$$

where $\Gamma(F_1, F_2)$ is the set of couplings of F_1 and F_2.

Alternative expressions:

$$\begin{aligned} W_1(F_1, F_2) &= \sup |\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \\ &= \int_{-\infty}^{\infty} |F_1(s) - F_2(s)| \mathrm{d}s = \int_0^1 |F_1^{-1}(\beta) - F_2^{-1}(\beta)| \mathrm{d}\beta, \end{aligned}$$

where the supremum is over all 1-Lipschitz functions $f : \mathbb{R} \to \mathbb{R}$.

Concentration of Mean-Variance

- Mean-variance (Markowitz, 1952) with risk tolerance γ:

$$\mathrm{MV} = \gamma\mu - \sigma^2.$$

Concentration of Mean-Variance

- Mean-variance (Markowitz, 1952) with risk tolerance γ:

$$\mathrm{MV} = \gamma\mu - \sigma^2.$$

- Empirical mean-variance

$$\widehat{\mathrm{MV}}_n := \gamma\hat{\mu}_n - \hat{\sigma}_n^2$$

where $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are the sample mean and sample variance respectively.

Concentration of Mean-Variance

- Mean-variance (Markowitz, 1952) with risk tolerance γ:

$$\mathrm{MV} = \gamma\mu - \sigma^2.$$

- Empirical mean-variance

$$\widehat{\mathrm{MV}}_n := \gamma\hat{\mu}_n - \hat{\sigma}_n^2$$

where $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ are the sample mean and sample variance respectively.

Lemma (Concentration bound for MV (simplified))

For any $\epsilon > 0$:

$$\Pr\left[|\widehat{\mathrm{MV}}_n - \mathrm{MV}| > \epsilon\right] \le 2\exp\left[-\frac{n\epsilon^2}{8\gamma^2\sigma^2}\right] + 2\exp\left(-\frac{n}{16}\min\left[\frac{\epsilon^2}{2\sigma^4}, \frac{\epsilon}{\sigma^2}\right]\right)$$

Concentration of Lipschitz-continuous Risk Measures

- Cassel et al. (2018) considered general risk measures that satisfy a Lipschitz requirement under some norm.

Concentration of Lipschitz-continuous Risk Measures

- Cassel et al. (2018) considered general risk measures that satisfy a Lipschitz requirement under some norm.
- Prashanth and Bhat (2020) use the Wasserstein distance as the underlying norm.

Definition

A risk measure $\rho(\cdot)$ is L-Lipschitz if for all cumulative distribution functions (F, G),

$$|\rho(F) - \rho(G)| \leq L\, W_1(F, G).$$

Idea: Use $\rho_n = \rho(F_n)$ as an estimate of $\rho(F) = \rho(X)$ ($X \sim F$), where

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\} \quad \text{for any } \; x \in \mathbb{R}.$$

Concentration of Lipschitz-continuous Risk Measures

Theorem

Let X be a sub-Gaussian r.v. with parameter σ^2. Suppose ρ an L-Lipschitz risk measure. Then, for every ϵ satisfying

$$\frac{256\sqrt{2}\sigma}{\sqrt{n}} < \frac{\epsilon}{L} < \frac{256\sqrt{2}\sigma}{\sqrt{n}} + 16\sigma\sqrt{2\mathrm{e}}, \quad \textit{i.e.,} \quad \epsilon = \Omega\Big(\frac{1}{\sqrt{n}}\Big)$$

we have

$$\Pr\left[|\rho_n - \rho(X)| > \epsilon\right] \le \exp\left(-\frac{n}{256\sigma^2\mathrm{e}}\Big(\frac{\epsilon}{L} - \frac{256\sqrt{2}\sigma}{\sqrt{n}}\Big)^2\right).$$

Concentration of CVaR

Definition

The Conditional Value-at-Risk (CVaR) at level $\alpha \in (0,1)$ for a r.v. X is

$$\mathrm{CVaR}_\alpha(X) := \inf_{\xi \in \mathbb{R}} \left\{ \xi + \frac{1}{(1-\alpha)} \mathbb{E}\left[(X-\xi)^+ \right] \right\}.$$

- Empirical CVaR given $\{X_i\}_{i=1}^n$:

$$c_{n,\alpha} = \inf_{\xi \in \mathbb{R}} \left\{ \xi + \frac{1}{n(1-\alpha)} \sum_{i=1}^n (X_i - \xi)^+ \right\}.$$

- But CVaR at level α is $\frac{1}{1-\alpha}$-Lipschitz

$$|\mathrm{CVaR}_\alpha(X) - \mathrm{CVaR}_\alpha(Y)| \le \frac{1}{1-\alpha} W_1(F_X, F_Y).$$

so we can use the preceding concentration bound.

Concentration of Spectral Risk Measures

Definition

Given a risk spectrum $\phi : [0,1] \to [0,\infty)$, the Spectral Risk Measure (SRM) M_ϕ associated with ϕ is defined by Acerbi (2002) as

$$M_\phi(X) = \int_0^1 \phi(\beta) F_X^{-1}(\beta)\, \mathrm{d}\beta.$$

- Suppose $\phi(u) \le K$ for all $u \in [0,1]$, then

$$|M_\phi(X) - M_\phi(Y)| \le K\, W_1(F_X, F_Y).$$

- Use the general concentration result for Lipschitz risk functionals and the estimator

$$m_{n,\phi} = \int_0^1 \phi(\beta) F_n^{-1}(\beta)\, \mathrm{d}\beta.$$

Application to UCB-type Bandit Algorithms

- K-armed bandit with unknown distributions $\nu = (\nu_1, \nu_2, \ldots, \nu_K)$.

Application to UCB-type Bandit Algorithms

- K-armed bandit with unknown distributions $\nu = (\nu_1, \nu_2, \ldots, \nu_K)$.
- At each time t, agent pulls an arm $A_t \in [K]$; this choice depends on the history $\mathcal{H}_{t-1} = (A_1, X_{1,A_1}, \ldots, A_{t-1}, X_{1,A_{t-1}})$.

Application to UCB-type Bandit Algorithms

- K-armed bandit with unknown distributions $\nu = (\nu_1, \nu_2, \ldots, \nu_K)$.
- At each time t, agent pulls an arm $A_t \in [K]$; this choice depends on the history $\mathcal{H}_{t-1} = (A_1, X_{1,A_1}, \ldots, A_{t-1}, X_{1,A_{t-1}})$.
- Seek to minimize the cumulative regret:

$$R_n^\rho(\nu, \pi) := \mathbb{E}\left[n \max_{1 \le i \le K} \rho(\nu_i) - \sum_{t=1}^{n} \rho(\nu_{A_t})\right],$$

Application to UCB-type Bandit Algorithms

- K-armed bandit with unknown distributions $\nu = (\nu_1, \nu_2, \ldots, \nu_K)$.
- At each time t, agent pulls an arm $A_t \in [K]$; this choice depends on the history $\mathcal{H}_{t-1} = (A_1, X_{1,A_1}, \ldots, A_{t-1}, X_{1,A_{t-1}})$.
- Seek to minimize the cumulative regret:

$$R_n^\rho(\nu, \pi) := \mathbb{E}\left[n \max_{1 \le i \le K} \rho(\nu_i) - \sum_{t=1}^{n} \rho(\nu_{A_t}) \right],$$

- Play all arms once, then

$$A_t = \arg\min_{1 \le i \le K} \mathrm{LCB}_t(i) \quad \text{where}$$

$$\mathrm{LCB}_t(i) = \rho_{i,T_i(t-1)} - w_{i,T_i(t-1)}$$

and $\rho_{i,T_i(t-1)}$ is the estimate of $\rho(\nu_i)$ with $T_i(t-1)$ samples.

Application to UCB-type Bandit Algorithms

Using the previous bounds for Lipschitz risk measures, we can obtain.

Theorem

The expected regret R_n^ρ of Risk-LCB satisfies the following bound:

$$R_n^\rho \leq \sum_{i:\Delta_i>0} \frac{4L^2\sigma^2[32\sqrt{\mathrm{e}\log n}+256\sqrt{2}]^2}{\Delta_i} + 5K\Delta_i$$

where

$$\Delta_i = \rho(\nu_{i^*}) - \rho(\nu_i).$$

Application to UCB-type Bandit Algorithms

Using the previous bounds for Lipschitz risk measures, we can obtain.

Theorem

The expected regret R_n^ρ of Risk-LCB satisfies the following bound:

$$R_n^\rho \le \sum_{i:\Delta_i>0} \frac{4L^2\sigma^2[32\sqrt{\mathrm{e}\log n}+256\sqrt{2}]^2}{\Delta_i} + 5K\Delta_i$$

where

$$\Delta_i = \rho(\nu_{i^*}) - \rho(\nu_i).$$

Bound mimics that of risk-neutral UCB except that Δ_i's depend on ρ.

Thompson Sampling-type Bandit Algorithms

For Gaussian bandits, Zhu and Tan (2020) considered MVTS with the following sampling and update strategy:

1. Sample precision $\tau_{i,t}$ from $\mathrm{Gamma}(\alpha_{i,t-1}, \beta_{i,t-1})$;
2. Sample $\theta_{i,t}$ from $\mathcal{N}(\hat{\mu}_{i,t-1}, 1/T_{i,t-1})$;
3. Play $A_t = \arg\max_{i\in[K]} \gamma\theta_{i,t} - 1/\tau_{i,t}$ and observe X_{t,A_t};
4. $\mathrm{Update}(\hat{\mu}_{A_t,t-1}, T_{A_t,t-1}, \alpha_{A_t,t-1}, \beta_{A_t,t-1})$ using Bayes rule.

Thompson Sampling-type Bandit Algorithms

For Gaussian bandits, Zhu and Tan (2020) considered MVTS with the following sampling and update strategy:

1. Sample precision $\tau_{i,t}$ from $\mathrm{Gamma}(\alpha_{i,t-1}, \beta_{i,t-1})$;
2. Sample $\theta_{i,t}$ from $\mathcal{N}(\hat{\mu}_{i,t-1}, 1/T_{i,t-1})$;
3. Play $A_t = \arg\max_{i\in[K]} \gamma\theta_{i,t} - 1/\tau_{i,t}$ and observe X_{t,A_t};
4. $\mathrm{Update}(\hat{\mu}_{A_t,t-1}, T_{A_t,t-1}, \alpha_{A_t,t-1}, \beta_{A_t,t-1})$ using Bayes rule.

Theorem (Zhu and Tan (2020))

The expected regret of MVTS is

$$\limsup_{n\to\infty} \frac{R_n^\rho}{\log n} \le \sum_{i=2}^{K} \max\left\{\frac{2}{\Gamma_{1,i}^2}, \frac{1}{h(\sigma_i^2/\sigma_1^2)}\right\}(\Delta_i + 2\overline{\Gamma}_i^2),$$

where $\Gamma_{1,j} := \mu_1 - \mu_j$, $\overline{\Gamma}_i^2 := \max_{j\in[K]}(\mu_i - \mu_i)^2$*,* $\Delta_i := \mathrm{MV}_{i^*} - \mathrm{MV}_i$*, and* $h(x) := \frac{1}{2}(x-1-\log x)$*. Bound is asymptotically optimal as* $\gamma \to \{0, \infty\}$*.*

Conclusion and Future Work

- Follow up work by Baudry et al. (2021) and Chang and Tan (2022) on Thompson sampling for CVaR and continuous risk measures

$$\limsup_{n\to\infty} \frac{R_n^\rho}{\log n} \le \sum_{i=2}^{K} \frac{\Delta_k^\rho}{K_{\inf}^\rho(\nu_k, r_1^\rho)} \quad \text{where} \quad K_{\inf}^\rho(\nu, r) = \inf_{\mu:\rho(\mu)\ge r} \mathrm{KL}(\mu, \nu).$$

Conclusion and Future Work

- Follow up work by Baudry et al. (2021) and Chang and Tan (2022) on Thompson sampling for CVaR and continuous risk measures

$$\limsup_{n\to\infty} \frac{R_n^\rho}{\log n} \le \sum_{i=2}^{K} \frac{\Delta_k^\rho}{K_{\inf}^\rho(\nu_k, r_1^\rho)} \quad \text{where} \quad K_{\inf}^\rho(\nu, r) = \inf_{\mu:\rho(\mu)\ge r} \mathrm{KL}(\mu, \nu).$$

- Many more Lipschitz risk measures, e.g., cumulative prospect theory (Jie et al., 2018; Prashanth et al., 2016) and utility-based shortfall risk (Artzner et al., 1999; Föllmer and Schied, 2002)

Conclusion and Future Work

- Follow up work by Baudry et al. (2021) and Chang and Tan (2022) on Thompson sampling for CVaR and continuous risk measures

$$\limsup_{n\to\infty} \frac{R_n^\rho}{\log n} \le \sum_{i=2}^{K} \frac{\Delta_k^\rho}{K_{\inf}^\rho(\nu_k, r_1^\rho)} \quad \text{where} \quad K_{\inf}^\rho(\nu, r) = \inf_{\mu:\rho(\mu)\ge r} \mathrm{KL}(\mu,\nu).$$

- Many more Lipschitz risk measures, e.g., cumulative prospect theory (Jie et al., 2018; Prashanth et al., 2016) and utility-based shortfall risk (Artzner et al., 1999; Föllmer and Schied, 2002)

- Best arm identification (pure exploration) problems under risk constraints
  - Fixed budget (Kagrecha et al., 2019; Prashanth et al., 2020; Zhang and Ong, 2021)
  - Fixed confidence (David and Shimkin, 2016; David et al., 2018; Hou et al., 2022; Szorenyi et al., 2015)

References I

C. Acerbi. Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7):1505–1518, 2002.

P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

D. Baudry, R. Gautron, E. Kaufmann, and O. Maillard. Optimal thompson sampling strategies for support-aware cvar bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 716–726. PMLR, Jul 2021.

A. Cassel, S. Mannor, and A. Zeevi. A general approach to multi-armed bandits under risk criteria. In *Proceedings of the 31st Conference On Learning Theory*, pages 1295–1306, 2018.

J. Q. L. Chang and V. Y. F. Tan. A Unifying Theory of Thompson Sampling for Continuous Risk-Averse Bandits. In *Proc. of the 36th AAAI Conference on Artificial Intelligence*. AAAI Press, Feb 2022.

Y. David and N. Shimkin. Pure exploration for max-quantile bandits. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 556–571. Springer, 2016.

Y. David, B. Szörényi, M. Ghavamzadeh, S. Mannor, and N. Shimkin. PAC bandits with risk constraints. In *ISAIM*, 2018.

H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447, 2002.

References II

Y. Hou, V. Y. F. Tan, and Z. Zhong. Almost optimal variance-constrained best arm identification, 2022. arXiv 2201.10142.

C. Jie, L. A. Prashanth, M. C. Fu, S. I. Marcus, and C. Szepesvári. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, 63(9): 2867–2882, 2018.

A. Kagrecha, J. Nair, and K. Jagannathan. Distribution oblivious, risk-aware algorithms for multi-armed bandits with unbounded rewards. In *Advances in Neural Information Processing Systems*, pages 11269–11278, 2019.

H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

L. A. Prashanth and S. P. Bhat. A Wasserstein distance approach for concentration of empirical risk estimates, 2020. arXiv 1902.10709v3.

L. A. Prashanth, J. Cheng, M. C. Fu, S. I. Marcus, and C. Szepesvári. Cumulative prospect theory meets reinforcement learning: prediction and control. In *International Conference on Machine Learning*, pages 1406–1415. PMLR, 2016.

L. A. Prashanth, K. Jagannathan, and R. K. Kolla. Concentration bounds for CVaR estimation: The cases of light-tailed and heavy-tailed distributions. In *International Conference on Machine Learning*, volume 119, pages 5577–5586. PMLR, 2020.

B. Szorenyi, R. Busa-Fekete, P. Weng, and E. Hüllermeier. Qualitative multi-armed bandits: A quantile-based approach. In *International Conference on Machine Learning*, pages 1660–1668. PMLR, 2015.

References III

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

M. Zhang and C. S. Ong. Quantile bandits for best arms identification. In *International Conference on Machine Learning*, pages 12513–12523. PMLR, 2021.

Q. Zhu and V. Y. F. Tan. Thompson sampling algorithms for mean-variance bandits. In *International Conference on Machine Learning*, pages 2645–2654. PMLR, 2020.