Local Metric Learning for Off-Policy Evaluation in Contextual Bandits with Continuous Actions

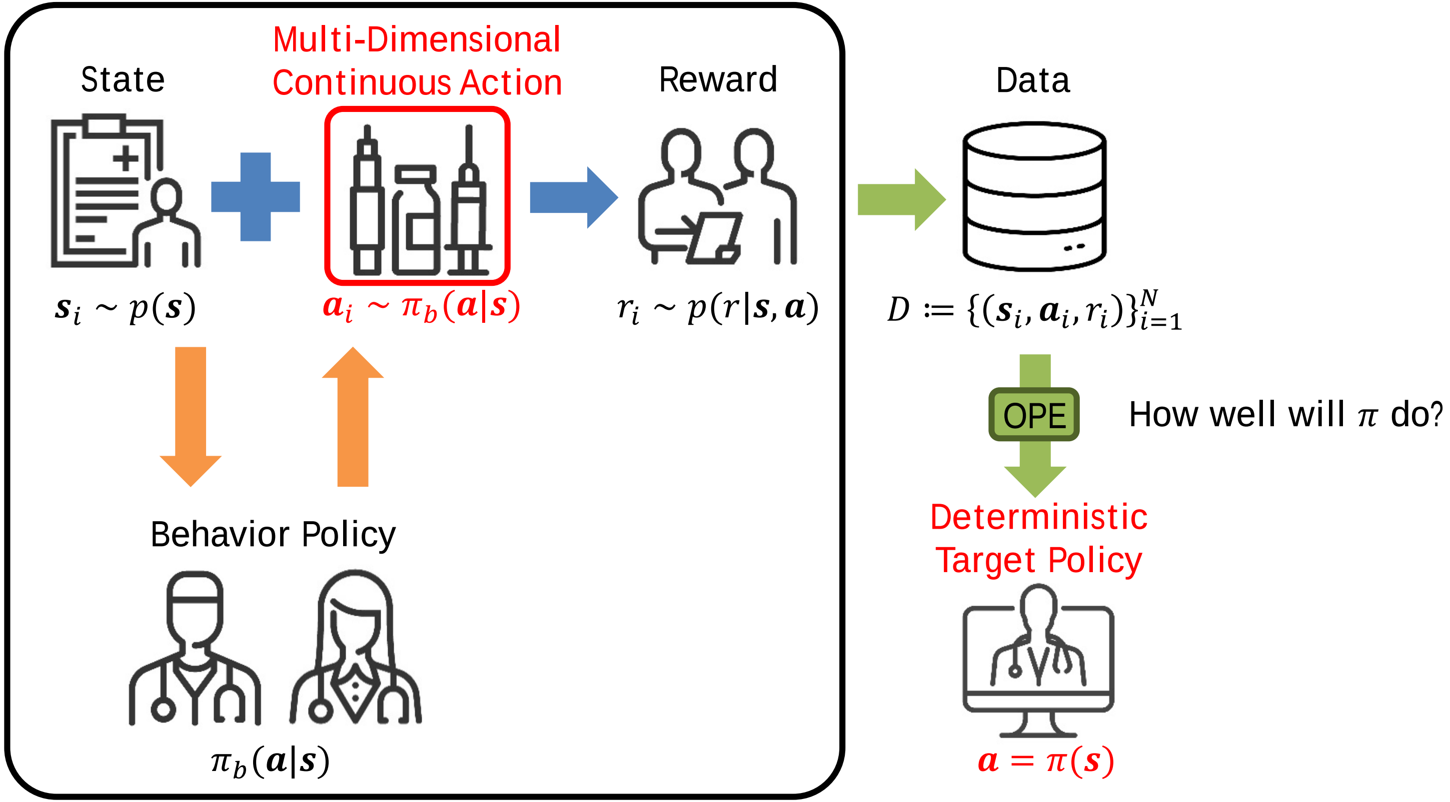
NeurIPS 2022

Haanvid Lee[1], Jongmin Lee[2], Yunseon Choi[1], Wonseok Jeon, Byung-Jun Lee[3,4], Yung-Kyun Noh[5,6], Kee-Eung Kim[1]

[1]KAIST, [2]UC Berkeley, [3]Korea Univ., [4]Gauss Labs Inc., [5]Hanyang Univ., [6]KIAS

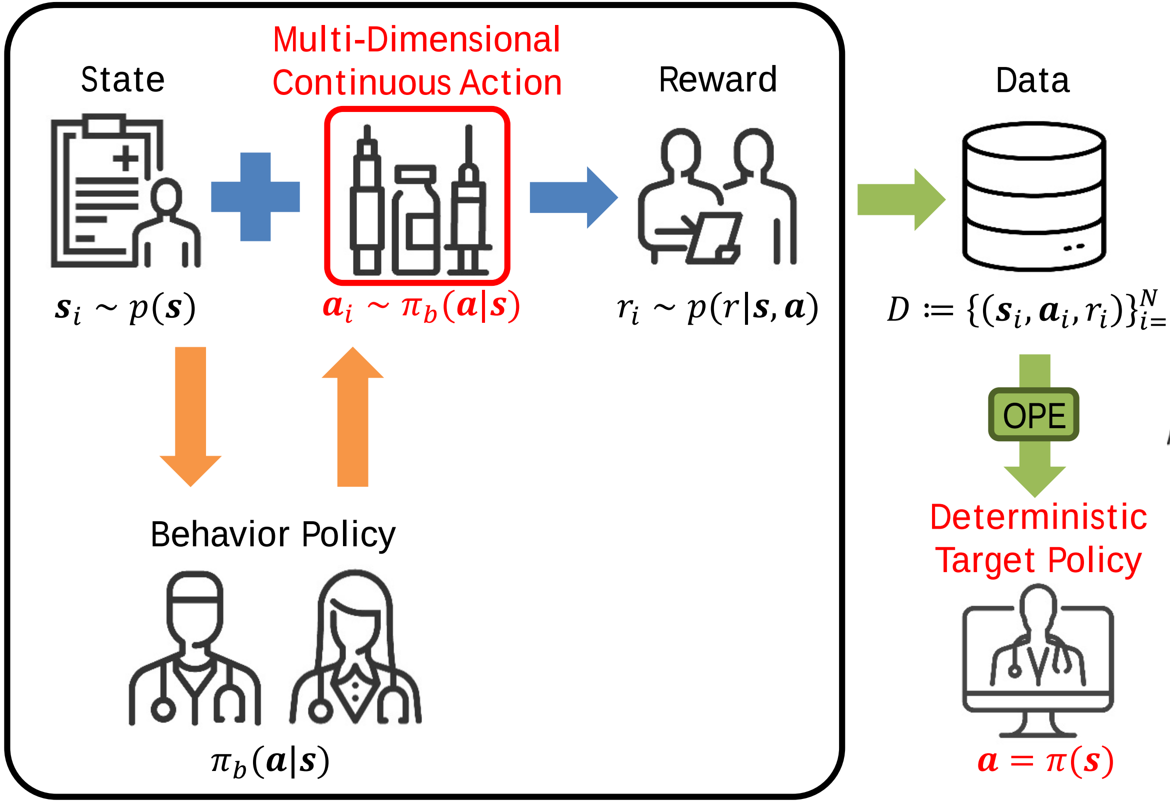
Off-Policy Evaluation (OPE) of Deterministic Policies

OPE: Evaluate a target policy using the data sampled by a behavior policy



Off-Policy Evaluation (OPE) of Deterministic Policies

OPE: Evaluate a target policy using the data sampled by a behavior policy



How well will π do?

$$\rho^\pi = \mathbb{E}_{s \sim p(s), a \sim \pi(a|s), r \sim p(r|s,a)}[r]$$

$$= \mathbb{E}_{s \sim p(s), a \sim \pi_b(a|s), r \sim p(r|s,a)}\left[\frac{\pi(a \mid s)}{\pi_b(a \mid s)} r\right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \frac{\delta\left(a_i - \pi\left(s_i\right)\right)}{\pi_b\left(a_i \mid s_i\right)} r_i$$

The importance sampling (IS) estimate is almost surely zero

Related Works

☐ Kernel-based methods

- Relax a deterministic target policy using a kernel

$$\rho^{\pi} \approx \frac{1}{N}\sum_{i=1}^{N}\frac{\delta\left(\boldsymbol{a}_i-\pi(\boldsymbol{s}_i)\right)}{\pi_b\left(\boldsymbol{a}_i \mid \boldsymbol{s}_i\right)} r_i$$

$$\approx \frac{1}{N h^{D_A}}\sum_{i=1}^{N} K\left(\frac{\boldsymbol{a}_i-\pi\left(\boldsymbol{s}_i\right)}{h}\right)\frac{r_i}{\pi_b\left(\boldsymbol{a}_i \mid \boldsymbol{s}_i\right)}$$

- Choose bandwidth h that best balances bias and variance
  - Select a bandwidth among a set of bandwidths using the Lepski's principle [1]
  - Choose the optimal bandwidth h^* that minimizes the leading-order MSE (LOMSE) [2]

$$\mathrm{LOMSE}\left(h, N, D_A\right) = \underbrace{h^4 C_b}_{(\text{leading-order bias})^2} + \underbrace{\frac{C_v}{N h^{D_A}}}_{(\text{leading-order variance})} :$$

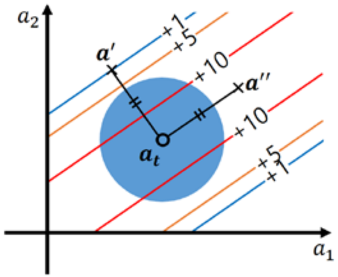
$$C_b := \frac{1}{4}\mathbb{E}_{\boldsymbol{s}\sim p(\boldsymbol{s})}\left[\nabla_{\boldsymbol{a}}^2 \mathbb{E}[r \mid \boldsymbol{s}, \boldsymbol{a}]\big|_{\boldsymbol{a}=\pi(\boldsymbol{s})}\right]^2,\ C_v := R(K)\mathbb{E}_{\boldsymbol{s}\sim p(\boldsymbol{s})}\left[\frac{\mathbb{E}\left[r^2 \mid \boldsymbol{s}, \boldsymbol{a}=\pi(\boldsymbol{s})\right]}{\pi_b(\boldsymbol{a}=\pi(\boldsymbol{s}) \mid \boldsymbol{s})}\right],\ R(K) := \int K(\boldsymbol{u})^2 d\boldsymbol{u}$$

[1] Yi Su et al. "Adaptive estimator selection for off-policy evaluation." ICML (2020)
[2] Nathan Kallus and Angela Zhou. "Policy evaluation and optimization with continuous treatments." AISTATS (2018)

Limitation of the Previous Works

☐ Usage of Euclidean distances induce excessive bias

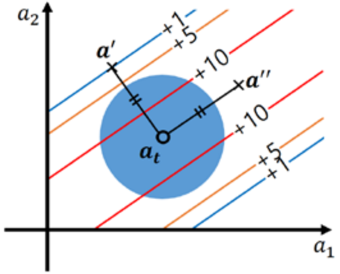
- Use Euclidean distances for measuring similarities between the target and behavior actions
- The Euclidean distance between actions may not reflect the similarity in the corresponding rewards
- Mahalanobis distance metric $A(\boldsymbol{s})$ locally learned at each state $\boldsymbol{s}$ can be used to reduce the bias



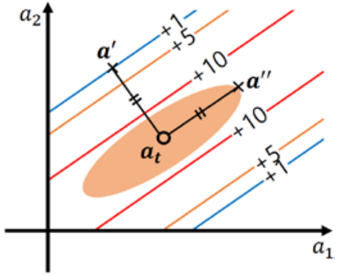
$$\frac{1}{2\pi}\exp\left(-\frac{(\boldsymbol{a}-\boldsymbol{a}_t)^\top(\boldsymbol{a}-\boldsymbol{a}_t)}{2}\right)$$

Limitation of the Previous Works

- ☐ Usage of Euclidean distances induce excessive bias
  - Use Euclidean distances for measuring similarities between the target and behavior actions
  - The Euclidean distance between actions may not reflect the similarity in the corresponding rewards
  - Mahalanobis distance metric $A(\boldsymbol{s})$ locally learned at each state $\boldsymbol{s}$ can be used to reduce the bias



$$\frac{1}{2\pi}\exp\left(-\frac{(\boldsymbol{a}-\boldsymbol{a}_t)^\top(\boldsymbol{a}-\boldsymbol{a}_t)}{2}\right)$$



$$\frac{1}{2\pi}\exp\left(-\frac{(\boldsymbol{a}-\boldsymbol{a}_t)^\top A(\boldsymbol{s})(\boldsymbol{a}-\boldsymbol{a}_t)}{2}\right)$$

$$\text{where } A(\boldsymbol{s}) \succ 0,\ A(\boldsymbol{s})^\top = A(\boldsymbol{s}),\ |A(\boldsymbol{s})| = 1,$$
$$\boldsymbol{a}_t = \pi(\boldsymbol{s})$$

- ☐ Could MSE be reduced by the reduction of bias with the metric?

Kernel Metric Learning for IS (KMIS)

☐ LOMSE has following characteristics when h^* is applied

- Bias of the isotropic kernel-based method dominates over the variance for $D_A \gg 4$ (Proposition 1)
- LOMSE approximates to C_b for a high action dimension $D_A \gg 4$ (Proposition 1)
- **For high dimensional action spaces, the MSE of a kernel-based IS estimator can be decreased by reducing C_b**
- Proposition 1 is adapted from [1]

For $D_A \gg 4$,

$$\underbrace{(h^*)^4 C_b}_{\text{(leading-order bias)}^2} \gg \underbrace{\frac{C_v}{N(h^*)^{D_A}}}_{\text{leading-order variance}} :$$

$$\text{LOMSE}\left(h^*, N, D_A\right) = N^{-\frac{4}{D_A+4}} \left(\left(\frac{D_A}{4}\right)^{\frac{4}{D_A+4}} + \left(\frac{4}{D_A}\right)^{\frac{D_A}{D_A+4}} \right) C_b^{\frac{D_A}{D_A+4}} C_v^{\frac{4}{D_A+4}} \approx C_b.$$

☐ Goal: Reduce C_b by applying $A(\boldsymbol{s})$

[1] Yung-Kyun Noh et al. "Generative local metric learning for kernel regression." NeurIPS (2017)

Kernel Metric Learning for IS (KMIS)

Derive the C_b in the Leading-Order Bias with a Metric

Upper Bound of $C_{b,A}$ as Minimization Objective

Compute the Closed-Form Solution

- ☐ C_b with a metric $A(\boldsymbol{s})$ (i.e. $C_{b,A}$)

$$C_{b,A} = \frac{1}{4}\mathbb{E}_{\boldsymbol{s}\sim p(\boldsymbol{s})}\left[\operatorname{tr}\left(A(\boldsymbol{s})^{-1}\mathbf{H}_{\boldsymbol{a}}\mathbb{E}[r \mid \boldsymbol{s}, \boldsymbol{a}]\big|_{\boldsymbol{a}=\pi(\boldsymbol{s})}\right)\right]^2$$

$\mathbf{H}_a$: Hessian operator

Kernel Metric Learning for IS (KMIS)

Derive the C_b in the Leading-Order Bias with a Metric

Upper Bound of $C_{b,A}$ as Minimization Objective

Compute the Closed-Form Solution

- ☐ C_b with a metric $A(\boldsymbol{s})$ (i.e. $C_{b,A}$)

$$C_{b,A} = \frac{1}{4}\mathbb{E}_{\boldsymbol{s}\sim p(\boldsymbol{s})}\left[\mathrm{tr}\left(A(\boldsymbol{s})^{-1}\mathbf{H}_{\boldsymbol{a}}\mathbb{E}[r \mid \boldsymbol{s},\boldsymbol{a}]\big|_{\boldsymbol{a}=\pi(\boldsymbol{s})}\right)\right]^2$$

$\mathbf{H}_a$: Hessian operator

- ☐ Minimize the upper bound of $C_{b,A}$ (i.e. $U_{b,A}$)

$$\min_{\substack{A:\ A(\boldsymbol{s})\succ 0,\\ A(\boldsymbol{s})=A(\boldsymbol{s})^\top, |A(\boldsymbol{s})|=1\ \forall \boldsymbol{s}}} U_{b,A} = \frac{1}{4}\mathbb{E}_{\boldsymbol{s}\sim p(\boldsymbol{s})}\left[\mathrm{tr}\left(A(\boldsymbol{s})^{-1}\mathbf{H}_{\boldsymbol{a}}\mathbb{E}[r \mid \boldsymbol{s},\boldsymbol{a}]\big|_{\boldsymbol{a}=\pi(\boldsymbol{s})}\right)^2\right]$$

Kernel Metric Learning for IS (KMIS)

Derive the C_b in the Leading-Order Bias with a Metric

Upper Bound of $C_{b,A}$ as Minimization Objective

Compute the Closed-Form Solution

- ☐ C_b with a metric $A(\boldsymbol{s})$ (i.e. $C_{b,A}$)

$$C_{b,A} = \frac{1}{4}\mathbb{E}_{\boldsymbol{s}\sim p(\boldsymbol{s})}\left[\mathrm{tr}\left(A(\boldsymbol{s})^{-1}\mathbf{H}_{\boldsymbol{a}}\mathbb{E}[r \mid \boldsymbol{s},\boldsymbol{a}]\big|_{\boldsymbol{a}=\pi(\boldsymbol{s})}\right)\right]^2$$

$\mathbf{H}_a$: Hessian operator

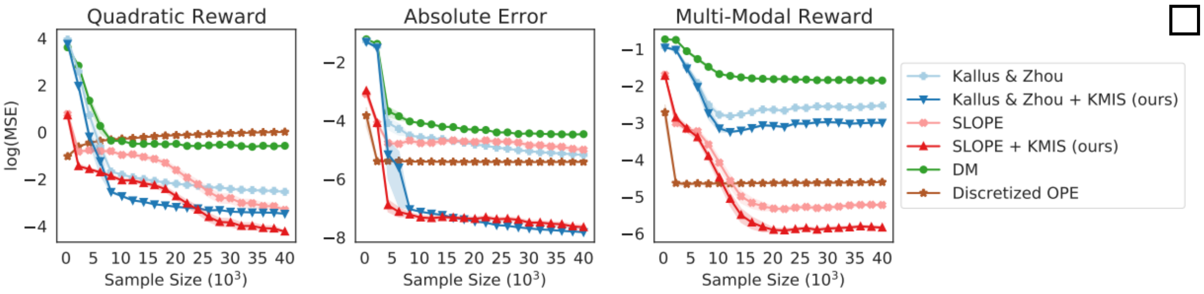
- ☐ Minimize the upper bound of $C_{b,A}$ (i.e. $U_{b,A}$)

$$\min_{\substack{A:\ A(\boldsymbol{s})\succ 0,\\ A(\boldsymbol{s})=A(\boldsymbol{s})^\top, |A(\boldsymbol{s})|=1\ \forall \boldsymbol{s}}} U_{b,A} = \frac{1}{4}\mathbb{E}_{\boldsymbol{s}\sim p(\boldsymbol{s})}\left[\mathrm{tr}\left(A(\boldsymbol{s})^{-1}\mathbf{H}_{\boldsymbol{a}}\mathbb{E}[r \mid \boldsymbol{s},\boldsymbol{a}]\big|_{\boldsymbol{a}=\pi(\boldsymbol{s})}\right)^2\right]$$

- ☐ Compute the closed-form metric matrix $A^*(\boldsymbol{s})$ that minimizes $U_{b,A}$ locally at each state $\boldsymbol{s}$ using the semi-definite programming solution from the work of Noh et al. [1] (Theorem 1)

[1] Yung-Kyun Noh et al. "Generative local metric learning for nearest neighbor classification." NeurIPS (2010)

Experiment: Synthetic Dataset



- Dataset ($\boldsymbol{s}, \boldsymbol{a} \in \mathbb{R}^2$)
  - Quadratic Reward

$$r \sim N\left(r(\boldsymbol{s}, \boldsymbol{a}), 0.5^2\right)$$

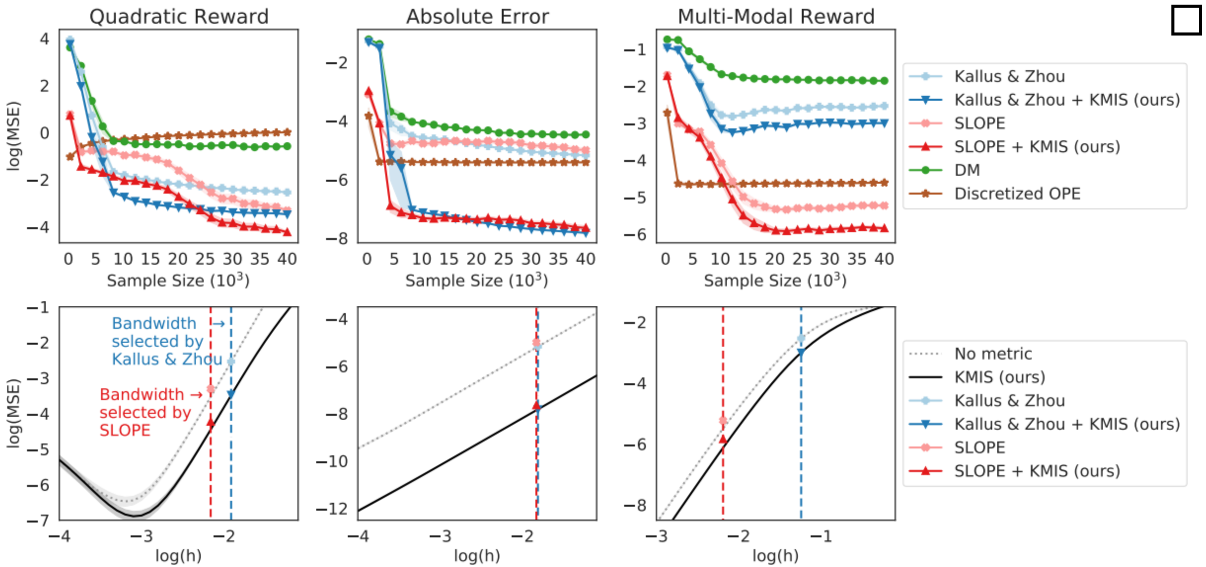
$$r(\boldsymbol{s}, \boldsymbol{a}) = -(\boldsymbol{s} - \boldsymbol{a})^\top \begin{bmatrix} 11 & 9 \\ 9 & 11 \end{bmatrix} (\boldsymbol{s} - \boldsymbol{a})$$

  - Absolute Error

$$r = -\left|0.5 s_1 - a_1\right|$$

  - Multi-Modal Reward
    - Multi-modal reward function composed of exponential functions and max operators

Experiment: Synthetic Dataset



☐ Dataset ($\boldsymbol{s}, \boldsymbol{a} \in \mathbb{R}^2$)

- Quadratic Reward

$$r \sim N\left(r(\boldsymbol{s}, \boldsymbol{a}), 0.5^2\right)$$

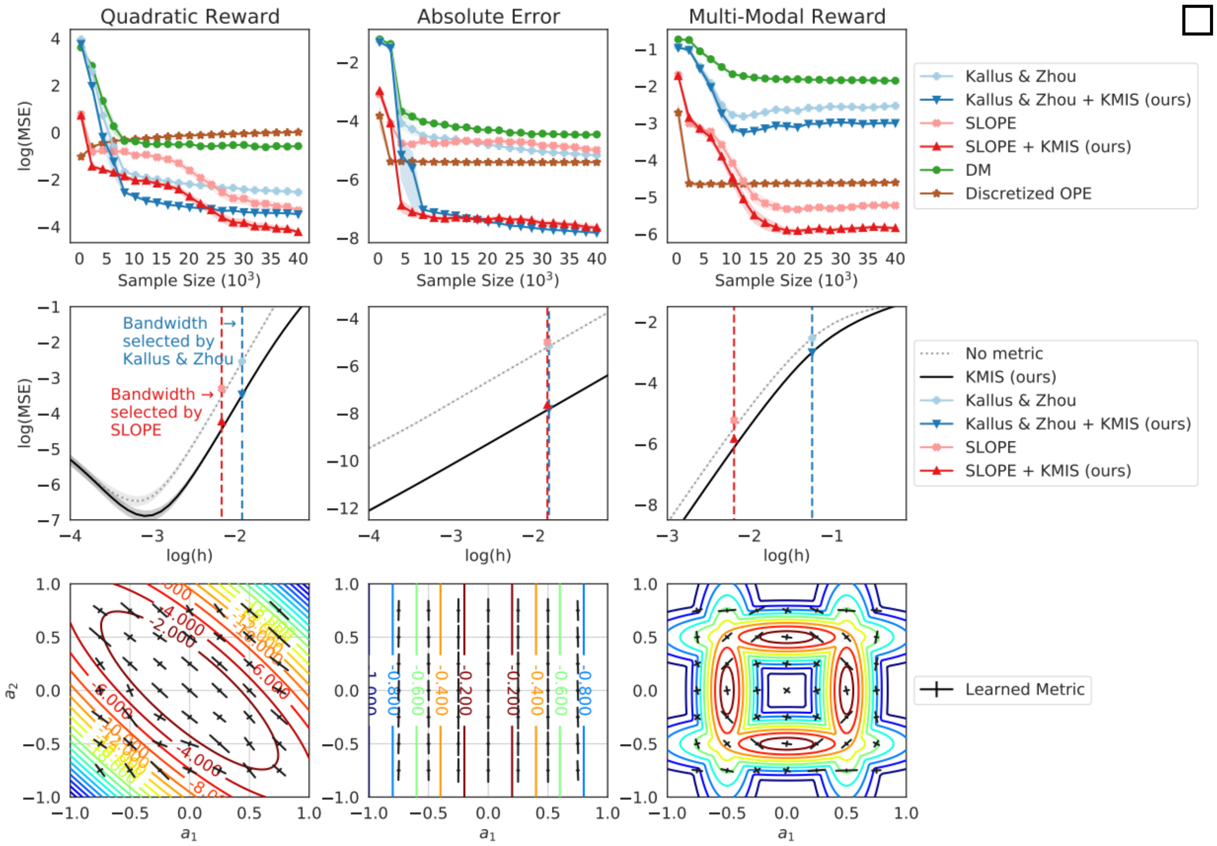
$$r(\boldsymbol{s}, \boldsymbol{a}) = -(\boldsymbol{s}-\boldsymbol{a})^\top \begin{bmatrix} 11 & 9 \\ 9 & 11 \end{bmatrix} (\boldsymbol{s}-\boldsymbol{a})$$

- Absolute Error

$$r = -\left|0.5 s_1 - a_1\right|$$

- Multi-Modal Reward
  - Multi-modal reward function composed of exponential functions and max operators

Experiment: Synthetic Dataset



- ☐ Dataset ($\boldsymbol{s}, \boldsymbol{a} \in \mathbb{R}^2$)
  - Quadratic Reward

    $$r \sim N\left(r(\boldsymbol{s}, \boldsymbol{a}), 0.5^2\right)$$

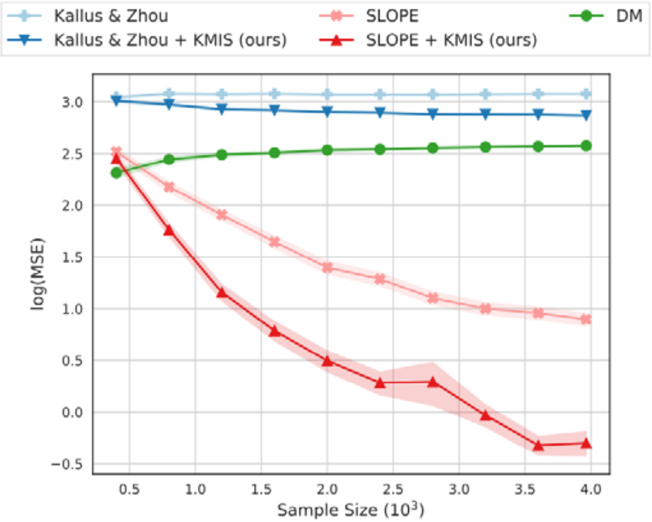
    $$r(\boldsymbol{s}, \boldsymbol{a}) = -(\boldsymbol{s} - \boldsymbol{a})^\top \begin{bmatrix} 11 & 9 \\ 9 & 11 \end{bmatrix} (\boldsymbol{s} - \boldsymbol{a})$$

  - Absolute Error

    $$r = -\left|0.5 s_1 - a_1\right|$$

  - Multi-Modal Reward
    - Multi-modal reward function composed of exponential functions and max operators

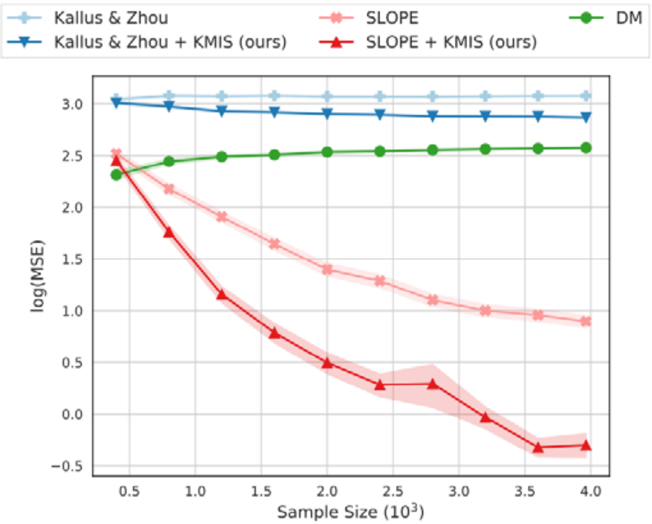
Experiment: Warfarin Dataset

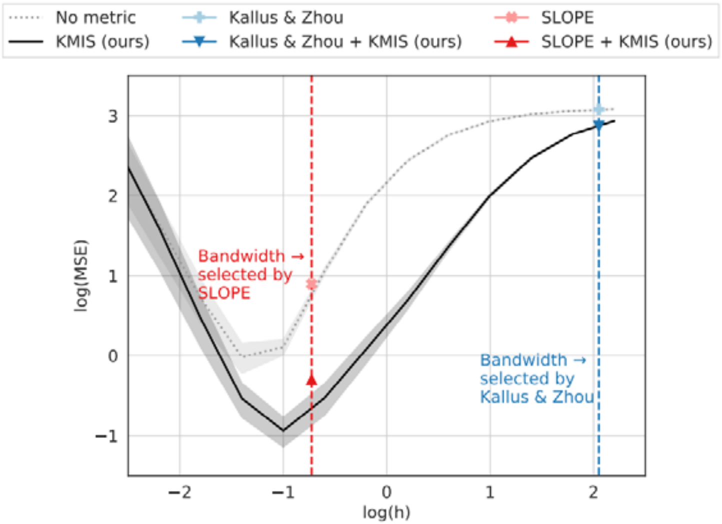


- Dataset
  - Warfarin dataset [1] contains patients' information and therapeutic doses
  - One dummy action dimension was added for testing the KMIS metric and the baselines

[1] International Warfarin Pharmacogenetics Consortium. "Estimation of the warfarin dose with clinical and pharmacogenetic data." New England Journal of Medicine 360.8 (2009): 753-764.

Experiment: Warfarin Dataset





- ☐ Dataset
  - Warfarin dataset [1] contains patients' information and therapeutic doses
  - One dummy action dimension was added for testing the KMIS metric and the baselines

[1] International Warfarin Pharmacogenetics Consortium. "Estimation of the warfarin dose with clinical and pharmacogenetic data." New England Journal of Medicine 360.8 (2009): 753-764.

Thank You

https://github.com/haanvid/kmis