# Kernelized Bandits Beyond Realizability

Davide Maran & Csaba Szepesvári

# Setting

Input space $\mathcal{X} = [-1, 1]^m$, function $f : \mathcal{X} \to \mathbb{R}$ to optimize.

# Setting

Input space $\mathcal{X} = [-1, 1]^m$, function $f : \mathcal{X} \to \mathbb{R}$ to optimize.

## Kernel

The kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is symmetric and positive semi-definite, and

$$\sup_{x,y \in \mathcal{X}} |k(x, y)| \leq \kappa < +\infty.$$

# Setting

Input space $\mathcal{X} = [-1, 1]^m$, function $f : \mathcal{X} \to \mathbb{R}$ to optimize.

## Kernel

The kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is symmetric and positive semi-definite, and

$$\sup_{x,y \in \mathcal{X}} |k(x, y)| \leq \kappa < +\infty.$$

## Interaction models

- *(Offline optimization)* Passive data:

$$(\boldsymbol{x}, \boldsymbol{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \text{ s.t. } y_t = f(x_t) + \eta_t$$

for independent $1-$subgaussian $\eta_t$.

# Setting

Input space $\mathcal{X} = [-1, 1]^m$, function $f : \mathcal{X} \to \mathbb{R}$ to optimize.

## Kernel

The kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is symmetric and positive semi-definite, and

$$\sup_{x,y \in \mathcal{X}} |k(x, y)| \leq \kappa < +\infty.$$

## Interaction models

- *(Offline optimization)* Passive data:

  $$(\boldsymbol{x}, \boldsymbol{y}) = ((x_1, y_1), \ldots, (x_n, y_n)) \text{ s.t. } y_t = f(x_t) + \eta_t$$

  for independent $1-$subgaussian $\eta_t$.
- *(Online optimization)* Choose sequentially $x_1, \ldots x_n$ and receive $y_t = f(x_t) + \eta_t$.

# Goals

## Target approximator

For $f_\star \in \mathcal{H} = \mathrm{RKHS}(k; \mathcal{X})$,

$$\boxed{\varepsilon := \sup_{x \in \mathcal{X}} |f(x) - f_\star(x)|}.$$

# Goals

## Target approximator

For $f_\star \in \mathcal{H} = \mathrm{RKHS}(k; \mathcal{X})$,

$$\boxed{\varepsilon := \sup_{x \in \mathcal{X}} |f(x) - f_\star(x)|}.$$

1. *Offline setting:* output $\widehat{x}_n$, minimize

$$r_n = f(x^*) - f(\widehat{x}_n).$$

# Goals

## Target approximator

For $f_\star \in \mathcal{H} = \mathrm{RKHS}(k; \mathcal{X})$,

$$\boxed{\varepsilon := \sup_{x \in \mathcal{X}} |f(x) - f_\star(x)|}.$$

1. *Offline setting:* output $\widehat{x}_n$, minimize

$$r_n = f(x^*) - f(\widehat{x}_n).$$

2. *Online setting:* minimize

$$R_n = \sum_{t=1}^{n} f(x^*) - f(x_t).$$

# Goals

## Target approximator

For $f_\star \in \mathcal{H} = \mathrm{RKHS}(k; \mathcal{X})$,

$$\boxed{\varepsilon := \sup_{x \in \mathcal{X}} |f(x) - f_\star(x)|}.$$

1. *Offline setting:* output $\widehat{x}_n$, minimize

$$r_n = f(x^*) - f(\widehat{x}_n).$$

2. *Online setting:* minimize

$$R_n = \sum_{t=1}^{n} f(x^*) - f(x_t).$$

In terms of $n, \varepsilon, \|f_\star\|_{\mathcal{H}}$ and properties of $\mathcal{H}$.

# Existing results [Bogunovic and Krause, 2021]

Regret bound:

$$R_n \leq \widetilde{O}\left(\gamma_n \sqrt{n} + \|f_\star\|_{\mathcal{H}} \sqrt{\gamma_n n} + \sqrt{\gamma}_n n\varepsilon\right).$$

# Existing results [Bogunovic and Krause, 2021]

Regret bound:

$$R_n \leq \widetilde{O}\left(\gamma_n \sqrt{n} + \|f_\star\|_{\mathcal{H}} \sqrt{\gamma_n n} + \sqrt{\gamma}_n n\varepsilon\right).$$

**Maximal information gain**

$\gamma_n := \max_{x_1,\ldots x_n \in \mathcal{X}} \log\det(I + \lambda^{-1}K)$, for $K_{ij} = k(x_i, x_j)$.

For $k$ in the Matérn family [Vakili et al., 2021]:

$$\gamma_n \asymp n^{\frac{m}{2\nu+m}},$$

except for log $-$factors.

# Existing results [Bogunovic and Krause, 2021]

Regret bound:

$$R_n \leq \widetilde{O}\left(\gamma_n \sqrt{n} + \|f_\star\|_{\mathcal{H}} \sqrt{\gamma_n n} + \sqrt{\gamma}_n n\varepsilon\right).$$

**Maximal information gain**

$\gamma_n := \max_{x_1,\dots x_n \in \mathcal{X}} \log\det(I + \lambda^{-1}K)$, for $K_{ij} = k(x_i, x_j)$.

For $k$ in the Matérn family [Vakili et al., 2021]:

$$\gamma_n \asymp n^{\frac{m}{2\nu+m}},$$

except for log −factors.

**Error amplification**

If $\varepsilon \approx n^{-\frac{m}{2\nu+m}}$ this gives *linear regret*.

# Proof backbone

## Notation

- $\mathrm{P}_{\boldsymbol{x},\lambda} g$: KRR estimator data $(\boldsymbol{x}, g(\boldsymbol{x}))$
- $\sigma_{\boldsymbol{x},\lambda}$: posterior standard deviation

$$\sigma_{\boldsymbol{x},\lambda}(x) := \sqrt{k(x,x) - \Psi_{\boldsymbol{x}}(x)^\top (K + \lambda I)^{-1} \Psi_{\boldsymbol{x}}(x)}.$$

# Proof backbone

## Notation

- $\mathrm{P}_{\boldsymbol{x},\lambda} g$: KRR estimator data $(\boldsymbol{x}, g(\boldsymbol{x}))$
- $\sigma_{\boldsymbol{x},\lambda}$: posterior standard deviation

$$\sigma_{\boldsymbol{x},\lambda}(x) := \sqrt{k(x,x) - \Psi_{\boldsymbol{x}}(x)^\top (K + \lambda I)^{-1} \Psi_{\boldsymbol{x}}(x)}.$$

$$|\mathrm{P}_{\boldsymbol{x},\lambda} f(x) - \mathrm{P}_{\boldsymbol{x},\lambda} f_\star(x)| \leq \sqrt{\frac{n}{\lambda}} \sigma_{\boldsymbol{x},\lambda}(x) \varepsilon.$$

# Proof backbone

## Notation

- $\mathrm{P}_{\boldsymbol{x},\lambda} g$: KRR estimator data $(\boldsymbol{x}, g(\boldsymbol{x}))$
- $\sigma_{\boldsymbol{x},\lambda}$: posterior standard deviation

$$\sigma_{\boldsymbol{x},\lambda}(x) := \sqrt{k(x,x) - \Psi_{\boldsymbol{x}}(x)^\top (K + \lambda I)^{-1} \Psi_{\boldsymbol{x}}(x)}.$$

$$|\mathrm{P}_{\boldsymbol{x},\lambda} f(x) - \mathrm{P}_{\boldsymbol{x},\lambda} f_\star(x)| \leq \sqrt{\frac{n}{\lambda}} \sigma_{\boldsymbol{x},\lambda}(x) \varepsilon.$$

Can we refine this result?

# Population version

$\boldsymbol{x} \overset{i.i.d.}{\sim} p$, $(\phi_i)_{i=1}^{\infty}$ orthogonal in $L^2(p)$ and

$$k(x, y) = \sum_{i=1}^{\infty} \mu_i \phi_i(x) \phi_i(y).$$

# Population version

$\boldsymbol{x} \overset{i.i.d.}{\sim} p$, $(\phi_i)_{i=1}^{\infty}$ orthogonal in $L^2(p)$ and

$$k(x, y) = \sum_{i=1}^{\infty} \mu_i \phi_i(x) \phi_i(y).$$

## Population operator

$$\mathrm{P}_\tau f(x) := \sum_{i=1}^{\infty} \frac{\mu_i f_i \phi_i(x)}{\tau + \mu_i} \qquad f_i = \langle f, \phi_i \rangle_{L^2(p)}.$$

For $\tau = \lambda/n$, $\|\mathrm{P}_\tau f - \mathrm{P}_{\boldsymbol{x},\lambda} f\|_\infty = o(1)$.

# Population version

$\boldsymbol{x} \overset{i.i.d.}{\sim} p$, $(\phi_i)_{i=1}^{\infty}$ orthogonal in $L^2(p)$ and

$$k(x, y) = \sum_{i=1}^{\infty} \mu_i \phi_i(x) \phi_i(y).$$

## Population operator

$$\mathrm{P}_\tau f(x) := \sum_{i=1}^{\infty} \frac{\mu_i f_i \phi_i(x)}{\tau + \mu_i} \qquad f_i = \langle f, \phi_i \rangle_{L^2(p)}.$$

For $\tau = \lambda/n$, $\|\mathrm{P}_\tau f - \mathrm{P}_{\boldsymbol{x},\lambda} f\|_\infty = o(1)$.

For bounded eigenfunctions

$$\|\mathrm{P}_\tau f - \mathrm{P}_\tau f_\star\|_\infty \leq \sqrt{d_{\mathrm{eff}}(k|\tau)}\varepsilon.$$

Note: $\sqrt{\frac{n}{\lambda}} \sup_{x \in \mathcal{X}} \sigma_{\boldsymbol{x},\lambda}(x) \propto \sqrt{d_{\mathrm{eff}}(k|\tau)} + o(1)$.

# Error decomposition

Idea:

$$r_n \leq 2 \cdot \sup_{x \in \mathcal{X}} |\mathrm{P}_{\boldsymbol{x},\lambda} \boldsymbol{y}(x) - f(x)|$$

# Error decomposition

Idea:

$$r_n \leq 2 \cdot \sup_{x \in \mathcal{X}} |\mathrm{P}_{\boldsymbol{x},\lambda} \boldsymbol{y}(x) - f(x)|$$

## Function decomposition

$f(x) = f_\star(x) + f_\varepsilon(x)$

# Error decomposition

Idea:

$$r_n \leq 2 \cdot \sup_{x \in \mathcal{X}} |\mathrm{P}_{\boldsymbol{x},\lambda} \boldsymbol{y}(x) - f(x)|$$

**Function decomposition**

$f(x) = f_\star(x) + f_\varepsilon(x)$

$$|\mathrm{P}_{\boldsymbol{x},\lambda} \boldsymbol{y}(x) - f(x)| \leq \underbrace{|\mathrm{P}_{\boldsymbol{x},\lambda} \boldsymbol{f}_\star(x) - f_\star(x)|}_{\text{stoch. y}} + \underbrace{|\mathrm{P}_{\boldsymbol{x},\lambda} f_\varepsilon(x) - \mathrm{P}_\tau f_\varepsilon(x)|}_{\text{stoch. design}}$$
$$+ \underbrace{|\mathrm{P}_\tau f_\varepsilon(x) - f_\varepsilon(x)|}_{\text{misspecification}}.$$

# Error decomposition

Idea:

$$r_n \leq 2 \cdot \sup_{x \in \mathcal{X}} |\mathsf{P}_{\boldsymbol{x},\lambda} \boldsymbol{y}(x) - f(x)|$$

## Function decomposition

$f(x) = f_\star(x) + f_\varepsilon(x)$

$$|\mathsf{P}_{\boldsymbol{x},\lambda} \boldsymbol{y}(x) - f(x)| \leq \underbrace{|\mathsf{P}_{\boldsymbol{x},\lambda} \boldsymbol{f_\star}(x) - f_\star(x)|}_{\text{stoch. y}} + \underbrace{|\mathsf{P}_{\boldsymbol{x},\lambda} f_\varepsilon(x) - \mathsf{P}_\tau f_\varepsilon(x)|}_{\text{stoch. design}}$$
$$+ \underbrace{|\mathsf{P}_\tau f_\varepsilon(x) - f_\varepsilon(x)|}_{\text{misspecification}}.$$

- stoch. y $\lesssim \sqrt{d_{\text{eff}}(k|\tau)} n^{-1/2} \|f_\star\|_{\mathcal{H}}$ [Hsu et al., 2012]
- stoch. design $\lesssim \kappa^2 n^{-1/2}(\varepsilon + n^{-1/2})$ [Smale and Zhou, 2005]
- misspecification $\lesssim \boxed{\Lambda(\mathsf{P}_\tau)\varepsilon}$

# Lebesgue constant analysis

## Lebesgue constant

For an operator $\Pi$, we call

$$\Lambda(\Pi) := \sup_{g \in L^\infty(\mathcal{X})} \frac{\|\Pi g\|_\infty}{\|g\|_\infty}.$$

# Lebesgue constant analysis

## Lebesgue constant

For an operator $\Pi$, we call

$$\Lambda(\Pi) := \sup_{g \in L^\infty(\mathcal{X})} \frac{\|\Pi g\|_\infty}{\|g\|_\infty}.$$

$L^\infty - L^1$ duality

$$\Lambda(\mathrm{P}_\tau) = \max_{x \in \mathcal{X}, f_i} \left\| \sum_{i=1}^{\infty} \frac{\mu_i f_i \phi_i(x)}{\tau + \mu_i} \phi_i(\cdot) \right\|_{L^1(p)}.$$

**Intractable due to the generality of the $\phi$'s**

# Spectral Lebesgue Growth

Many different kernels use the same features and different eigenvalues!

# Spectral Lebesgue Growth

Many different kernels use the same features and different eigenvalues!

**Bochner's theorem**

Stationary periodic kernels admit a discrete Fourier basis as their exact eigenfunctions.

# Spectral Lebesgue Growth

Many different kernels use the same features and different eigenvalues!

**Bochner's theorem**

Stationary periodic kernels admit a discrete Fourier basis as their exact eigenfunctions.

Fourier features $\implies$ S.L.G.

**S.L.G.**

$$\sup_{x \in \mathcal{X}} \left\| \sum_{j=1}^{i} \phi_j(x)\phi_j(\cdot) \right\|_{L^1(p)} \lesssim \log(e+i).$$

# The question turns on eigenvalue profiles

Call $\mu_i$ the eigenvalues sorted according to S.L.G.

# The question turns on eigenvalue profiles

Call $\mu_i$ the eigenvalues sorted according to S.L.G.

**Typical case = worst case**

For a random eigenvalue profile $\mu_i \sim i^{-s} X_i$ with $X_i \overset{i.i.d.}{\sim} \text{Bern}(1/2)$,

$$\Lambda(\mathrm{P}_\tau) \asymp \sqrt{d_{\text{eff}}(k|\tau)} \qquad a.s.$$

# The question turns on eigenvalue profiles

Call $\mu_i$ the eigenvalues sorted according to S.L.G.

**Typical case = worst case**

For a random eigenvalue profile $\mu_i \sim i^{-s} X_i$ with $X_i \overset{i.i.d.}{\sim} \mathrm{Bern}(1/2)$,

$$\Lambda(\mathsf{P}_\tau) \asymp \sqrt{d_{\mathrm{eff}}(k|\tau)} \qquad a.s.$$

**Monotone spectrum**

If $\mu_i \downarrow 0$ (monotonically),

$$\Lambda(\mathsf{P}_\tau) \lesssim \log(e + \kappa/\tau).$$

No effective dimension, only log-dependence in $\kappa$ and $1/\tau$.

# The question turns on eigenvalue profiles

Call $\mu_i$ the eigenvalues sorted according to S.L.G.

## Typical case = worst case

For a random eigenvalue profile $\mu_i \sim i^{-s} X_i$ with $X_i \overset{i.i.d.}{\sim} \text{Bern}(1/2)$,

$$\Lambda(\mathsf{P}_\tau) \asymp \sqrt{d_{\text{eff}}(k|\tau)} \qquad a.s.$$

## Monotone spectrum

If $\mu_i \downarrow 0$ (monotonically),

$$\Lambda(\mathsf{P}_\tau) \lesssim \log(e + \kappa/\tau).$$

No effective dimension, only log-dependence in $\kappa$ and $1/\tau$.

## Tensorized kernels

$k(x, y) = \prod_{j=1}^{m} k_j(x_j, y_j)$ where satisfies S.L.G. and $\mu_i \downarrow 0$.

$$\Lambda(\mathsf{P}_\tau) \lesssim \log(e + \kappa^m/\tau)^{2m-1}.$$

# Online optimization

Controlling the regret

$$R_n = \sum_{t=1}^{n} f(x^*) - f(x_t)$$

requires *exploration-exploitation trade-off*.

# Online optimization

Controlling the regret

$$R_n = \sum_{t=1}^{n} f(x^*) - f(x_t)$$

requires *exploration-exploitation trade-off*.

## Explore-then-commit

1. Sample uniformly the first $n_0 = n^\alpha$ queries
2. Put the remaining $n - n^\alpha$ on $\widehat{x}_{n_0}$

# Online optimization

Controlling the regret

$$R_n = \sum_{t=1}^{n} f(x^*) - f(x_t)$$

requires *exploration-exploitation trade-off*.

## Explore-then-commit

1. Sample uniformly the first $n_0 = n^\alpha$ queries
2. Put the remaining $n - n^\alpha$ on $\widehat{x}_{n_0}$
3. **Sub-optimal**

# Online optimization

Controlling the regret

$$R_n = \sum_{t=1}^{n} f(x^*) - f(x_t)$$

requires *exploration-exploitation trade-off*.

## Explore-then-commit

1. Sample uniformly the first $n_0 = n^\alpha$ queries
2. Put the remaining $n - n^\alpha$ on $\widehat{x}_{n_0}$
3. **Sub-optimal**

## Smooth level curves

1. Assume level sets $\{f(\cdot) \geq M\}$ are smooth

# Online optimization

Controlling the regret

$$R_n = \sum_{t=1}^{n} f(x^*) - f(x_t)$$

requires *exploration-exploitation trade-off*.

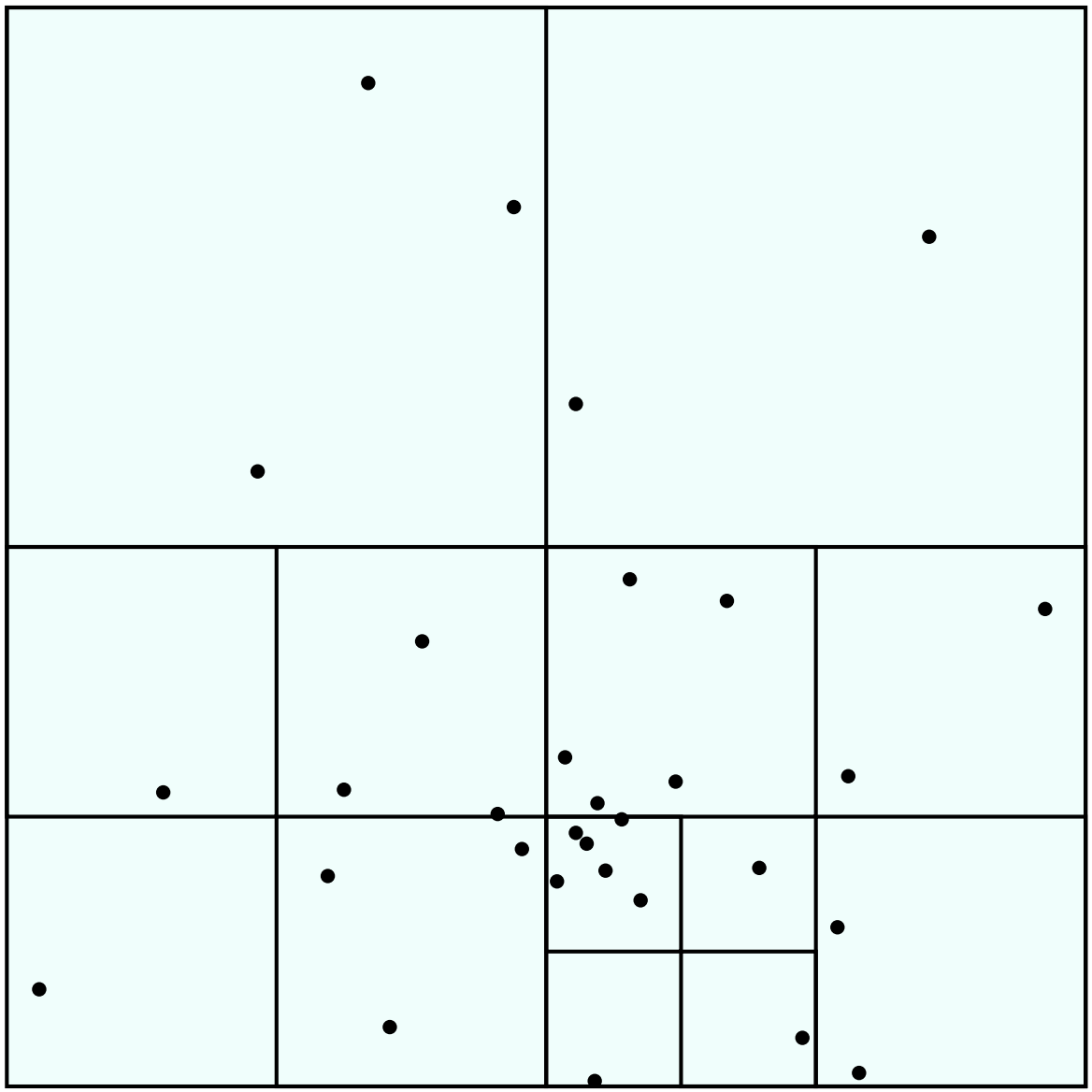
## Explore-then-commit

1. Sample uniformly the first $n_0 = n^\alpha$ queries
2. Put the remaining $n - n^\alpha$ on $\widehat{x}_{n_0}$
3. **Sub-optimal**

## Smooth level curves

1. Assume level sets $\{f(\cdot) \geq M\}$ are smooth
2. **Very restrictive**

# $\pi$−GP-UCB [Janz et al., 2020]



Split any region $A$ with too many point into $2^m$ sub-regions. Always keeps that

$$\gamma_{|A \cap \boldsymbol{x}|} \leq \text{polylog}(n).$$

# Key observations

## Domain splitting

Let $\Pi : L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X})$ such that

$$\Pi[f;\mathcal{X}](x) = \underbrace{1\{x \in \mathcal{X}_0\}\Pi[f;\mathcal{X}_0](x)}_{\Pi_0} + \underbrace{1\{x \in \mathcal{X}_0^c\}\Pi[f;\mathcal{X}_0^c](x)}_{\Pi_1},$$

# Key observations

## Domain splitting

Let $\Pi : L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X})$ such that

$$\Pi[f;\mathcal{X}](x) = \underbrace{1\{x \in \mathcal{X}_0\}\Pi[f;\mathcal{X}_0](x)}_{\Pi_0} + \underbrace{1\{x \in \mathcal{X}_0^c\}\Pi[f;\mathcal{X}_0^c](x)}_{\Pi_1},$$

Then, $\Lambda(\Pi) = \max\{\Lambda(\Pi_0), \Lambda(\Pi_1)\}$.

# Key observations

## Domain splitting

Let $\Pi : L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X})$ such that

$$\Pi[f;\mathcal{X}](x) = \underbrace{1\{x \in \mathcal{X}_0\}\Pi[f;\mathcal{X}_0](x)}_{\Pi_0} + \underbrace{1\{x \in \mathcal{X}_0^c\}\Pi[f;\mathcal{X}_0^c](x)}_{\Pi_1},$$

Then, $\Lambda(\Pi) = \max\{\Lambda(\Pi_0), \Lambda(\Pi_1)\}$.

## Lebesgue constant and information gain

For KRR, $\Lambda(\mathrm{P}_{A\cap \boldsymbol{x},\lambda}) \lesssim \sqrt{\gamma_{|A\cap \boldsymbol{x}|}}$.

# Key observations

## Domain splitting

Let $\Pi : L^\infty(\mathcal{X}) \to L^\infty(\mathcal{X})$ such that

$$\Pi[f;\mathcal{X}](x) = \underbrace{1\{x \in \mathcal{X}_0\}\Pi[f;\mathcal{X}_0](x)}_{\Pi_0} + \underbrace{1\{x \in \mathcal{X}_0^c\}\Pi[f;\mathcal{X}_0^c](x)}_{\Pi_1},$$

Then, $\Lambda(\Pi) = \max\{\Lambda(\Pi_0), \Lambda(\Pi_1)\}$.

## Lebesgue constant and information gain

For KRR, $\Lambda(\mathrm{P}_{A\cap \boldsymbol{x},\lambda}) \lesssim \sqrt{\gamma_{|A\cap \boldsymbol{x}|}}$. With $\pi-\mathrm{GP\text{-}UCB}$,

$$\gamma_{|A\cap \boldsymbol{x}|} \leq \mathrm{polylog}(n),$$

on **all regions**.

# Final regret bound

## Eigendecay on subdomains

Restricting $k$ to an hypercube of side $\rho > 0$ gives $\mu_{\rho,i} \leq \rho^{\alpha}\mu_i$ and bounded $\phi_{\rho,i}$.

# Final regret bound

**Eigendecay on subdomains**

Restricting $k$ to an hypercube of side $\rho > 0$ gives $\mu_{\rho,i} \leq \rho^{\alpha} \mu_i$ and bounded $\phi_{\rho,i}$.

$k \in$ Matérn family $\implies \alpha = 2\nu$.

# Final regret bound

**Eigendecay on subdomains**

Restricting $k$ to an hypercube of side $\rho > 0$ gives $\mu_{\rho,i} \leq \rho^{\alpha} \mu_i$ and bounded $\phi_{\rho,i}$.

$k \in$ Matérn family $\implies \alpha = 2\nu$.

$$\boxed{R_n \leq \widetilde{O}\left(\|f_\star\|_{\mathcal{H}} \sqrt{\gamma_n n} + n\varepsilon\right)}.$$

# Open questions

- Is there some useful kernel such that the $\sqrt{d_{\text{eff}}(k|\tau)}\varepsilon$ is inevitable?

# Open questions

- Is there some useful kernel such that the $\sqrt{d_{\text{eff}}(k|\tau)}\varepsilon$ is inevitable?
- Offline method $\approx$ localization of the spectrum, online method $\approx$ localization of the domain

# Bibliography


Bogunovic, I. and Krause, A. (2021).
Misspecified gaussian process bandit optimization.
*Advances in neural information processing systems*, 34:3004–3015.

Hsu, D., Kakade, S. M., and Zhang, T. (2012).
Random design analysis of ridge regression.
In *Conference on learning theory*, pages 9.1–9.24. JMLR Workshop and Conference Proceedings.

Janz, D., Burt, D., and González, J. (2020).
Bandit optimisation of functions in the matérn kernel rkhs.
In *International Conference on Artificial Intelligence and Statistics*, pages 2486–2495. PMLR.

Smale, S. and Zhou, D.-X. (2005).
Shannon sampling ii: Connections to learning theory.
*Applied and Computational Harmonic Analysis*, 19(3):285–302.

Vakili, S., Khezeli, K., and Picheny, V. (2021).
On information gain and regret bounds in gaussian process bandits.
In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR.