Vector Balancing and Kernel Density Estimation

Rainie Bozzai

Joint Work with Thomas Rothvoss

June 12, 2024

The Vector Balancing Problem

Given:

- Symmetric convex bodies
  $K, Q \subset \mathbb{R}^d$

The Vector Balancing Problem

Given:

- Symmetric convex bodies $K, Q \subset \mathbb{R}^d$
- Associated Minkowski norms $\| \cdot \|_K, \| \cdot \|_Q$

The Vector Balancing Problem

Given:

- Symmetric convex bodies $K, Q \subset \mathbb{R}^d$
- Associated Minkowski norms $\|\cdot\|_K, \|\cdot\|_Q$
- Vectors $v_1, ..., v_n \in K$

The Vector Balancing Problem

Given:

- Symmetric convex bodies $K, Q \subset \mathbb{R}^d$
- Associated Minkowski norms $\| \cdot \|_K, \| \cdot \|_Q$
- Vectors $v_1, ..., v_n \in K$

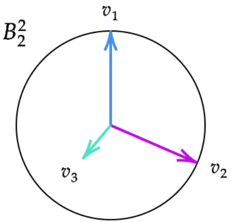
Goal: find signs $\varepsilon \in \{\pm 1\}^n$ to minimize

$$\Big\| \sum_{i \in [n]} \varepsilon_i v_i \Big\|_Q$$

The Vector Balancing Problem

Given:

- Symmetric convex bodies $K, Q \subset \mathbb{R}^d$
- Associated Minkowski norms $\| \cdot \|_K, \| \cdot \|_Q$
- Vectors $v_1, ..., v_n \in K$



Goal: find signs $\varepsilon \in \{\pm 1\}^n$ to minimize

$$\Big\| \sum_{i \in [n]} \varepsilon_i v_i \Big\|_Q$$

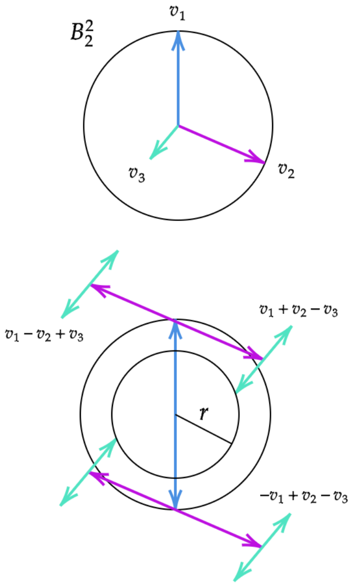
The Vector Balancing Problem

Given:

- Symmetric convex bodies $K, Q \subset \mathbb{R}^d$
- Associated Minkowski norms $\|\cdot\|_K, \|\cdot\|_Q$
- Vectors $v_1, \ldots, v_n \in K$

Goal: find signs $\varepsilon \in \{\pm 1\}^n$ to minimize

$$\Big\| \sum_{i \in [n]} \varepsilon_i v_i \Big\|_Q$$



Discrepancy Theory

$$\mathrm{vb}(K, Q) := \sup\left\{ \min_{x \in \{\pm 1\}^n} \left\| \sum_{i=1}^{n} x_i v_i \right\|_Q : \; n \in \mathbb{N}, \; v_1, \ldots, v_n \in K \right\}$$

Discrepancy Theory

$$\mathrm{vb}(K, Q) := \sup \left\{ \min_{x \in \{\pm 1\}^n} \left\| \sum_{i=1}^{n} x_i v_i \right\|_Q : \ n \in \mathbb{N}, \ v_1, ..., v_n \in K \right\}$$

Linear Algebraic Version: For $A \in \mathbb{R}^{d \times n}$, the *discrepancy* of A is

$$\mathrm{disc}(A) = \min_{x \in \{\pm 1\}^n} \|Ax\|_\infty$$

Discrepancy Theory

$$\mathrm{vb}(K, Q) := \sup \left\{ \min_{x \in \{\pm 1\}^n} \left\| \sum_{i=1}^{n} x_i v_i \right\|_Q : \; n \in \mathbb{N}, \; v_1, \ldots, v_n \in K \right\}$$

Linear Algebraic Version: For $A \in \mathbb{R}^{d \times n}$, the *discrepancy* of A is

$$\mathrm{disc}(A) = \min_{x \in \{\pm 1\}^n} \|Ax\|_\infty$$

Why discrepancy?

For $A \in \{0, 1\}^n$ we recover the *combinatorial discrepancy problem*.

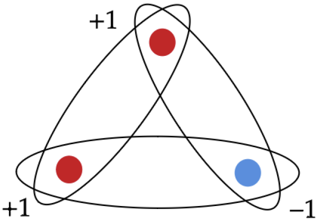
Discrepancy Theory

$$\mathrm{vb}(K, Q) := \sup \left\{ \min_{x \in \{\pm 1\}^n} \left\| \sum_{i=1}^{n} x_i v_i \right\|_Q : \; n \in \mathbb{N}, \; v_1, \ldots, v_n \in K \right\}$$

Linear Algebraic Version: For $A \in \mathbb{R}^{d \times n}$, the *discrepancy* of A is

$$\mathrm{disc}(A) = \min_{x \in \{\pm 1\}^n} \|Ax\|_\infty$$

Why discrepancy?
For $A \in \{0,1\}^n$ we recover the *combinatorial discrepancy problem*.



Kernel Density Estimation

Definition

Let ρ be a probability distribution on $\mathcal{D}$, $\{X_1, ..., X_n\} \sim \rho$ i.i.d., and $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$. The *Kernel Density Estimator* (KDE) given by $\mathcal{K}$ is then

$$KDE_X(y) = \frac{1}{n} \sum_{i=1}^{n} K(X_i, y).$$

Kernel Density Estimation

Definition

Let ρ be a probability distribution on $\mathcal{D}$, $\{X_1, ..., X_n\} \sim \rho$ i.i.d., and $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$. The *Kernel Density Estimator* (KDE) given by $\mathcal{K}$ is then

$$KDE_X(y) = \frac{1}{n} \sum_{i=1}^{n} K(X_i, y).$$

Example: Gaussian kernel, $K_G(x, y) = \exp(-\alpha^2 \|x - y\|_2^2)$

Kernel Density Estimation

Definition

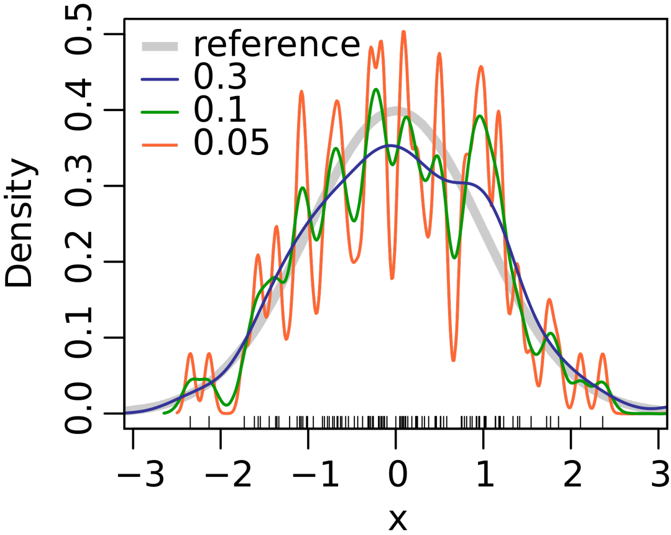
Let ρ be a probability distribution on $\mathcal{D}$, $\{X_1, ..., X_n\} \sim \rho$ i.i.d., and $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$. The *Kernel Density Estimator* (KDE) given by $\mathcal{K}$ is then

$$KDE_X(y) = \frac{1}{n} \sum_{i=1}^{n} K(X_i, y).$$

Example: Gaussian kernel, $K_G(x, y) = \exp(-\alpha^2 \|x - y\|_2^2)$

- Well-known that $KDE_X(y)$ approximates ρ at the minimax optimal rate as $|X| \to \infty$ for "well-behaved" kernels

Kernel Density Estimation



Coresets for KDEs

Definition (ε-Coreset)

Given $\varepsilon > 0$, $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, and a data set $X \subseteq \mathcal{D}$, an ε-coreset for K is a subset $Q \subseteq X$ such that

$$\|\mathrm{KDE}_X - \mathrm{KDE}_Q\|_\infty = \sup_{y \in \mathcal{D}} \left| \frac{1}{|X|} \sum_{x \in X} K(x, y) - \frac{1}{|Q|} \sum_{q \in Q} K(q, y) \right| \leq \varepsilon.$$

Coresets for KDEs

Definition (ε-Coreset)

Given $\varepsilon > 0$, $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, and a data set $X \subseteq \mathcal{D}$, an ε-coreset for K is a subset $Q \subseteq X$ such that

$$\|\mathrm{KDE}_X - \mathrm{KDE}_Q\|_\infty = \sup_{y \in \mathcal{D}} \left| \frac{1}{|X|} \sum_{x \in X} K(x, y) - \frac{1}{|Q|} \sum_{q \in Q} K(q, y) \right| \leq \varepsilon.$$

- The *coreset complexity* of a kernel function K is the minimum size coreset given any choice of $X \subseteq \mathcal{D}$

Coresets for KDEs

Definition (ε-Coreset)

Given $\varepsilon > 0$, $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, and a data set $X \subseteq \mathcal{D}$, an ε-coreset for K is a subset $Q \subseteq X$ such that

$$\|\mathrm{KDE}_X - \mathrm{KDE}_Q\|_\infty = \sup_{y \in \mathcal{D}} \left| \frac{1}{|X|} \sum_{x \in X} K(x, y) - \frac{1}{|Q|} \sum_{q \in Q} K(q, y) \right| \le \varepsilon.$$

- The *coreset complexity* of a kernel function K is the minimum size coreset given any choice of $X \subseteq \mathcal{D}$
- Bounds depend on K and the dimension d of the data, and are independent of the size and choice of X.

Kernel Discrepancy

Fix a kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [-1, 1]$ and data set $X \subseteq \mathcal{D}$.

Kernel Discrepancy

Fix a kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [-1, 1]$ and data set $X \subseteq \mathcal{D}$. Assume that $|\mathcal{D}| = d < \infty$, $|X| = n$ (think $d \gg n$).

Kernel Discrepancy

Fix a kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [-1, 1]$ and data set $X \subseteq \mathcal{D}$. Assume that $|\mathcal{D}| = d < \infty$, $|X| = n$ (think $d \gg n$).

$$K \in \mathbb{R}^{d \times n}, \quad K_{ij} = K(y_i, x_j); \text{ columns } K^x \in \mathbb{R}^d \text{ indexed by } X$$

Kernel Discrepancy

Fix a kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [-1, 1]$ and data set $X \subseteq \mathcal{D}$. Assume that $|\mathcal{D}| = d < \infty$, $|X| = n$ (think $d \gg n$).

$$K \in \mathbb{R}^{d \times n}, \quad K_{ij} = K(y_i, x_j); \text{ columns } K^x \in \mathbb{R}^d \text{ indexed by } X$$

Suppose we can find **balanced** signs $\varepsilon \in \{\pm 1\}^n$ so that

$$\left\| \sum_{x \in X} \varepsilon_x K^x \right\|_\infty = \sup_{i \in [d]} \left| \sum_{x \in X} \varepsilon_x K(y_i, x) \right| \leq f(n, d).$$

The Halving Trick

We take $S_- := \{x : \varepsilon_x = -1\}$ to be our coreset. For any $y \in \mathcal{D}$:

The Halving Trick

We take $S_- := \{x : \varepsilon_x = -1\}$ to be our coreset. For any $y \in \mathcal{D}$:

$$\begin{aligned}
|\mathrm{KDE}_X(y) - \mathrm{KDE}_{S_-}(y)| &= \left| \tfrac{1}{|X|} \sum_{x \in X} K(x,y) - \tfrac{1}{|X|/2} \sum_{x \in S_-} K(x,y) \right| \\
&= \frac{1}{|X|} \left| \sum_{x \in X} K(x,y) - 2 \sum_{x \in S_-} K(x,y) \right| \\
&= \frac{1}{|X|} \left| \sum_{x \in S_+} K(x,y) - \sum_{x \in S_-} K(x,y) \right| \\
&= \frac{1}{|X|} \left| \sum_{x \in X} \varepsilon_x K(x,y) \right| \\
&\leq f(n,d)/n.
\end{aligned}$$

The Halving Trick

We take $S_- := \{x : \varepsilon_x = -1\}$ to be our coreset. For any $y \in \mathcal{D}$:

$$
\begin{aligned}
|\mathrm{KDE}_X(y) - \mathrm{KDE}_{S_-}(y)| &= \left| \tfrac{1}{|X|} \sum_{x \in X} K(x,y) - \tfrac{1}{|X|/2} \sum_{x \in S_-} K(x,y) \right| \\
&= \frac{1}{|X|} \left| \sum_{x \in X} K(x,y) - 2 \sum_{x \in S_-} K(x,y) \right| \\
&= \frac{1}{|X|} \left| \sum_{x \in S_+} K(x,y) - \sum_{x \in S_-} K(x,y) \right| \\
&= \frac{1}{|X|} \left| \sum_{x \in X} \varepsilon_x K(x,y) \right| \\
&\leq f(n,d)/n.
\end{aligned}
$$

Hope: $f(n,d)$ decays polynomially with n, get bound via iteration.

Our Problem

Goal: Given a kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [0, 1]$, $X \subseteq \mathcal{D}$, bound $\mathrm{disc}(K)$.

Our Problem

Goal: Given a kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [0, 1]$, $X \subseteq \mathcal{D}$, bound $\mathrm{disc}(K)$.

- Phillips, 2013: $\mathrm{disc}(K) = O(n^{1/2-1/d}\sqrt{\log(n)})$ for *Lipchitz*, rotation-and-shift invariant kernels; $\Theta(1)$ for $d = 1$

Our Problem

Goal: Given a kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [0, 1]$, $X \subseteq \mathcal{D}$, bound $\mathrm{disc}(K)$.

- Phillips, 2013: $\mathrm{disc}(K) = O(n^{1/2-1/d}\sqrt{\log(n)})$ for *Lipchitz*, rotation-and-shift invariant kernels; $\Theta(1)$ for $d = 1$
- Phillips and Tai, 2020: $\mathrm{disc}(K) = O(\sqrt{d \log n})$ for *positive definite*, Lipchitz, decaying kernels; $\Omega(\sqrt{d})$ with added assumption of rotation, shift invariance

Our Problem

Goal: Given a kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [0, 1]$, $X \subseteq \mathcal{D}$, bound $\mathrm{disc}(K)$.

- Phillips, 2013: $\mathrm{disc}(K) = O(n^{1/2-1/d}\sqrt{\log(n)})$ for *Lipchitz*, rotation-and-shift invariant kernels; $\Theta(1)$ for $d = 1$
- Phillips and Tai, 2020: $\mathrm{disc}(K) = O(\sqrt{d \log n})$ for *positive definite*, Lipchitz, decaying kernels; $\Omega(\sqrt{d})$ with added assumption of rotation, shift invariance
- Tai, 2022: $\mathrm{disc}(K_G) = O(1)$ for d constant

Our Problem

Goal: Given a kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [0,1]$, $X \subseteq \mathcal{D}$, bound $\mathrm{disc}(K)$.

- Phillips, 2013: $\mathrm{disc}(K) = O(n^{1/2-1/d}\sqrt{\log(n)})$ for *Lipchitz*, rotation-and-shift invariant kernels; $\Theta(1)$ for $d = 1$
- Phillips and Tai, 2020: $\mathrm{disc}(K) = O(\sqrt{d \log n})$ for *positive definite*, Lipchitz, decaying kernels; $\Omega(\sqrt{d})$ with added assumption of rotation, shift invariance
- Tai, 2022: $\mathrm{disc}(K_G) = O(1)$ for d constant

And a related result...

- Karnin and Liberty, 2019: $\mathrm{disc}(K) = O(\sqrt{d})$ under **very** strong assumptions on $\mathcal{D}$

Our Contributions

Broad Contribution: Ability to account for the geometry of the data set X by dropping the discretization of $\mathcal{D}$.

Our Contributions

Broad Contribution: Ability to account for the geometry of the data set X by dropping the discretization of $\mathcal{D}$.

- $\mathrm{disc}(K_G), \mathrm{disc}(K_L) = O(\sqrt{d \log \log n})$ for uniformly bounded datasets X.

Our Contributions

Broad Contribution: Ability to account for the geometry of the data set X by dropping the discretization of $\mathcal{D}$.

- $\mathrm{disc}(K_G), \mathrm{disc}(K_L) = O(\sqrt{d \log \log n})$ for uniformly bounded datasets X.

- $\mathrm{disc}(K_L) = O(\sqrt{\log \log n})$ for d constant.

Our Contributions

Broad Contribution: Ability to account for the geometry of the data set X by dropping the discretization of $\mathcal{D}$.

- $\mathrm{disc}(K_G), \mathrm{disc}(K_L) = O(\sqrt{d \log \log n})$ for uniformly bounded datasets X.
- $\mathrm{disc}(K_L) = O(\sqrt{\log \log n})$ for d constant.
- $\mathrm{disc}(K) = \sqrt{d \log(2 \max\{\alpha, 1\})}$ for the exponential, JS, and Hellinger kernels

Our Contributions

Broad Contribution: Ability to account for the geometry of the data set X by dropping the discretization of $\mathcal{D}$.

- $\mathrm{disc}(K_G), \mathrm{disc}(K_L) = O(\sqrt{d \log \log n})$ for uniformly bounded datasets X.
- $\mathrm{disc}(K_L) = O(\sqrt{\log \log n})$ for d constant.
- $\mathrm{disc}(K) = \sqrt{d \log(2 \max\{\alpha, 1\})}$ for the exponential, JS, and Hellinger kernels
  - Significantly improved dependence on bandwidth parameter α for the exponential kernel

Kernel Properties

For our purposes: the kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [-1, 1]$, with $\mathcal{D} \subseteq \mathbb{R}^d$, is

Kernel Properties

For our purposes: the kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [-1, 1]$, with $\mathcal{D} \subseteq \mathbb{R}^d$, is

- **Positive definite**: $\forall x_1, ..., x_m \in \mathcal{D}$, the Gram matrix $G_{ij} = K(x_i, x_j)$ is positive definite.

Kernel Properties

For our purposes: the kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [-1, 1]$, with $\mathcal{D} \subseteq \mathbb{R}^d$, is

- **Positive definite**: $\forall x_1, ..., x_m \in \mathcal{D}$, the Gram matrix $G_{ij} = K(x_i, x_j)$ is positive definite.
- Normalized: $\forall x \in \mathcal{D}$, $K(x, x) = 1$.

Kernel Properties

For our purposes: the kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [-1, 1]$, with $\mathcal{D} \subseteq \mathbb{R}^d$, is

- **Positive definite**: $\forall x_1, ..., x_m \in \mathcal{D}$, the Gram matrix $G_{ij} = K(x_i, x_j)$ is positive definite.
- Normalized: $\forall x \in \mathcal{D}$, $K(x, x) = 1$.
- (Generally) of the form $K(x, y) = \kappa(\alpha \|x - y\|_2)$, $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to [-1, 1]$ strictly decreasing and continuous

Kernel Properties

For our purposes: the kernel $\mathcal{K} : \mathcal{D} \times \mathcal{D} \to [-1, 1]$, with $\mathcal{D} \subseteq \mathbb{R}^d$, is

- **Positive definite**: $\forall x_1, ..., x_m \in \mathcal{D}$, the Gram matrix $G_{ij} = K(x_i, x_j)$ is positive definite.
- Normalized: $\forall x \in \mathcal{D}$, $K(x, x) = 1$.
- (Generally) of the form $K(x, y) = \kappa(\alpha \|x - y\|_2)$, $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to [-1, 1]$ strictly decreasing and continuous

Examples:

- Gaussian: $K_G(x, y) = \exp(-\alpha^2 \|x - y\|_2^2)$, $x, y \in \mathbb{R}^d$
- Laplacian: $K_L(x, y) = \exp(-\alpha \|x - y\|_2)$, $x, y \in \mathbb{R}^d$
- JS: $K_{JS}(x, y) = \exp\big(-\alpha\big(H(\frac{x+y}{2}) - \frac{H(x)+H(y)}{2}\big)\big)$, $x, y \in \Delta^d$

Key Theorem in Discrepancy Method

Theorem (Banaszczyk, '98)

Given any convex body $K \subseteq \mathbb{R}^m$ *of Gaussian measure* $\gamma_m(K) \geq 1/2$*, and vector* $v_1, ..., v_n \in B_2^d$*, there exist signs* $\varepsilon \in \{\pm 1\}^n$ *such that* $\sum_{i \in [n]} \varepsilon_i v_i \in CK$*,* $C > 0$ *an absolute constant.*

Key Theorem in Discrepancy Method

Theorem (Banaszczyk, '98)

Given any convex body $K \subseteq \mathbb{R}^m$ *of Gaussian measure* $\gamma_m(K) \geq 1/2$*, and vector* $v_1, ..., v_n \in B_2^d$*, there exist signs* $\varepsilon \in \{\pm 1\}^n$ *such that* $\sum_{i \in [n]} \varepsilon_i v_i \in CK$, $C > 0$ *an absolute constant.*

Theorem (Dadush et.al., 2018)

There is a polynomial-time randomized algorithm that takes as input vectors $v_1, ..., v_n \in \mathbb{R}^m$ *of* ℓ_2 *norm at most* 1 *and outputs random signs* $\varepsilon \in \{\pm 1\}^n$ *such that the (mean-zero) random variable* $\sum_{i \in [n]} \varepsilon_i v_i$ *is* $O(1)$*-subgaussian.*

Reproducing Kernel Hilbert Spaces

Theorem (Moore-Aronszajn, 1950)

Let T be a set and K a positive definite function on $T \times T$. Then there is a map $\phi : T \to \mathcal{H}_K$ to a unique corresponding reproducing kernel Hilbert space $\mathcal{H}_K$ so that for any $s, t \in T$,

$$K(s,t) = \langle \phi(s), \phi(t) \rangle_{\mathcal{H}_K}.$$

Reproducing Kernel Hilbert Spaces

Theorem (Moore-Aronszajn, 1950)

Let T be a set and K a positive definite function on $T \times T$. Then there is a map $\phi : T \to \mathcal{H}_K$ to a unique corresponding reproducing kernel Hilbert space $\mathcal{H}_K$ so that for any $s, t \in T$,

$$K(s, t) = \langle \phi(s), \phi(t) \rangle_{\mathcal{H}_K}.$$

For fixed kernel K with associated RKHS map $\phi : \mathcal{D} \to \mathcal{H}_K$, take the collection of vectors $\{\phi(x)\})_{x \in X}$.

Reproducing Kernel Hilbert Spaces

Theorem (Moore-Aronszajn, 1950)

Let T be a set and K a positive definite function on $T \times T$. Then there is a map $\phi : T \to \mathcal{H}_K$ to a unique corresponding reproducing kernel Hilbert space $\mathcal{H}_K$ so that for any $s, t \in T$,

$$K(s,t) = \langle \phi(s), \phi(t) \rangle_{\mathcal{H}_K}.$$

For fixed kernel K with associated RKHS map $\phi : \mathcal{D} \to \mathcal{H}_K$, take the collection of vectors $\{\phi(x)\})_{x \in X}$.

- Dadush et. al $\implies$ $\exists \varepsilon \in \{\pm 1\}^n$ such that $\sum_{x \in X} \varepsilon_x \phi(x)$ is $O(1)$-subgaussian.

Our Approach

Restated Goal: bound

$$\sup_{y \in Q} |\langle \Sigma, \phi(y) \rangle_{\mathcal{H}_K}| = \sup_{y \in Q} \Big| \sum_{x \in X} K(x, y) \Big|,$$

where $\Sigma := \sum_{x \in X} \varepsilon_x \phi(x)$ is $O(1)$-subgaussian.

Our Approach

Restated Goal: bound

$$\sup_{y \in Q} |\langle \Sigma, \phi(y) \rangle_{\mathcal{H}_K}| = \sup_{y \in Q} \left| \sum_{x \in X} K(x, y) \right|,$$

where $\Sigma := \sum_{x \in X} \varepsilon_x \phi(x)$ is $O(1)$-subgaussian.

Theorem (Dudley's Integral Inequality)

Let $(X_t)_{t \in T}$ be a mean zero random process on a pseudometric space (T, d) satisfying $\|X_t - X_s\|_{\psi_2} \leq d(t, s)$ for all $t, s \in T$. Then

$$\mathbb{E} \sup_{t \in T} X_t \lesssim \int_0^{\mathrm{diam}(d)} \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon.$$

Our Approach

Restated Goal: bound

$$\sup_{y \in Q} |\langle \Sigma, \phi(y) \rangle_{\mathcal{H}_K}| = \sup_{y \in Q} \Big| \sum_{x \in X} K(x, y) \Big|,$$

where $\Sigma := \sum_{x \in X} \varepsilon_x \phi(x)$ is $O(1)$-subgaussian.

Theorem (Dudley's Integral Inequality)

Let $(X_t)_{t \in T}$ be a mean zero random process on a pseudometric space (T, d) satisfying $\|X_t - X_s\|_{\psi_2} \leq d(t, s)$ for all $t, s \in T$. Then

$$\mathbb{E} \sup_{t \in T} X_t \lesssim \int_0^{\mathrm{diam}(d)} \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon.$$

Key Idea: Apply Dudley to $\Sigma_y := \langle \Sigma, \phi(y) \rangle$ for $y \in Q$ with d given by $\| \cdot \|_{\mathcal{H}_K}$, the *kernel distance*

Future Ideas

Key Question: How can we account for the geometry of the data in applications to get better bounds?

Future Ideas

Key Question: How can we account for the geometry of the data in applications to get better bounds?

- In general, we expect the bandwidth parameter to depend on n. If we account for this in our iteration, can we get better bounds?

Future Ideas

Key Question: How can we account for the geometry of the data in applications to get better bounds?

- In general, we expect the bandwidth parameter to depend on n. If we account for this in our iteration, can we get better bounds?
- Can assumed properties of the distribution give us better bounds?

Thank You!

Questions? :)