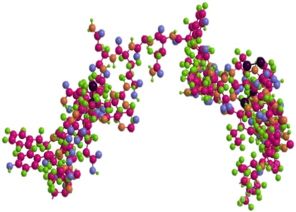
Learning under latent symmetries

Subhro Ghosh
National University of Singapore

The problem of Cryo Electron Microscopy

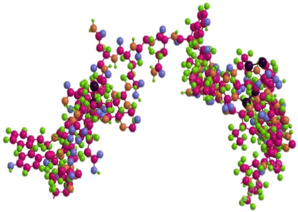
The Nobel Prize in Chemistry 2017 was awarded to Jacques Dubochet, Joachim Frank and Richard Henderson "for developing cryo-electron microscopy for the high- resolution structure determination of biomolecules in solution".

The problem of Cryo Electron Microscopy



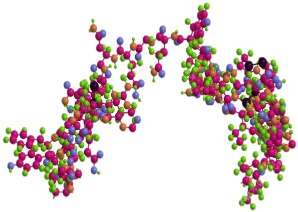
- Cryo-EM is an imaging technique for for the high-resolution structure determination of molecules.

The problem of Cryo Electron Microscopy



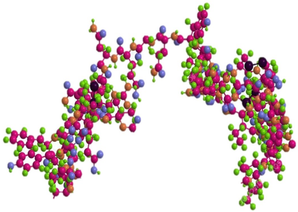
- Cryo-EM is an imaging technique for for the high-resolution structure determination of molecules.
- Each measurement consists of a noisy image of an unknown molecule
- The molecule is rotated by an unknown rotation in SO(3) in each measurement.
- The task is then to reconstruct the molecule density from many such measurements.

The problem of Cryo Electron Microscopy



- The reconstruction problem in Cryo-EM has received significant attention from the computational perspective.
- Statistical properties remain largely unexplored.

The problem of Cryo Electron Microscopy



- The reconstruction problem in Cryo-EM has received significant attention from the computational perspective.
- Statistical properties remain largely unexplored.
- Key features as a stochastic model :
  - The latent group action in each observation — in this case, a rotation
  - The presence of extremely high levels of noise

The stochastic model

The orbit recovery problem

- Objective : To determine $\theta^* \in \mathbb{R}^p$

The stochastic model

The orbit recovery problem

- Objective : To determine $\theta^* \in \mathbb{R}^p$
- Observations : $Y_i = G_i \cdot \theta^* + \xi_i; i = 1, 2, \ldots, n,$

The stochastic model

The orbit recovery problem

- Objective : To determine $\theta^* \in \mathbb{R}^p$
- Observations : $Y_i = G_i \cdot \theta^* + \xi_i; i = 1, 2, \dots, n,$ where
  - G_i are i.i.d. uniform according to Haar measure on a compact subgroup $\mathcal{G} \subset O(p)$

The stochastic model

The orbit recovery problem

- Objective : To determine $\theta^* \in \mathbb{R}^p$
- Observations : $Y_i = G_i \cdot \theta^* + \xi_i; i = 1, 2, \ldots, n$, where
  - G_i are i.i.d. uniform according to Haar measure on a compact subgroup $\mathcal{G} \subset O(p)$
  - ξ_i are i.i.d. standard Gaussians $\sim N_p(0, \sigma^2 I_p)$.

The stochastic model

The orbit recovery problem

- Objective : To determine $\theta^* \in \mathbb{R}^p$
- Observations : $Y_i = G_i \cdot \theta^* + \xi_i; i = 1, 2, \ldots, n$, where
  - G_i are i.i.d. uniform according to Haar measure on a compact subgroup $\mathcal{G} \subset O(p)$
  - ξ_i are i.i.d. standard Gaussians $\sim N_p(0, \sigma^2 I_p)$.

Observe : We can only recover θ^* up to its orbit under the action of $\mathcal{G}$; in other words, we can only hope to find the set

$$\mathcal{O}_{\theta^*} := \{\theta \in \mathbb{R}^p : \theta = g \cdot \theta^* \text{ for some } g \in \mathcal{G}\}.$$

The stochastic model

The orbit recovery problem : special cases

- Learning a bag of numbers : $\theta^* \in \mathbb{R}^p, \mathcal{G} = S_p \subset O(p)$

The stochastic model

The orbit recovery problem : special cases

- Learning a bag of numbers : $\theta^* \in \mathbb{R}^p, \mathcal{G} = S_p \subset O(p)$
- Learning a rigid body : $\theta^* \in \mathbb{R}^{k \times N}, \mathcal{G} = SO(k)$, acting diagonally on the columns of $\mathbb{R}^{k \times N}$

The stochastic model

The orbit recovery problem : special cases

- Learning a bag of numbers : $\theta^* \in \mathbb{R}^p, \mathcal{G} = S_p \subset O(p)$
- Learning a rigid body : $\theta^* \in \mathbb{R}^{k \times N}, \mathcal{G} = SO(k)$, acting diagonally on the columns of $\mathbb{R}^{k \times N}$
- Multi Reference Alignment (MRA) : $\theta^* \in \mathbb{R}^p, \mathcal{G} = \mathbb{Z}/p\mathbb{Z}$, acting as cyclic shifts on the coordinates of $\mathbb{R}^p$

The stochastic model

The orbit recovery problem : special cases

- Learning a bag of numbers : $\theta^* \in \mathbb{R}^p, \mathcal{G} = S_p \subset O(p)$
- Learning a rigid body : $\theta^* \in \mathbb{R}^{k \times N}, \mathcal{G} = SO(k)$, acting diagonally on the columns of $\mathbb{R}^{k \times N}$
- Multi Reference Alignment (MRA) : $\theta^* \in \mathbb{R}^p, \mathcal{G} = \mathbb{Z}/p\mathbb{Z}$, acting as cyclic shifts on the coordinates of $\mathbb{R}^p$
- Spherical registration problem : Learn $f : \mathbb{S}^2 \to \mathbb{R}$ from noisy measurements of $f(g^{-1} \bullet)$ with $g \in SO(3)$

The stochastic model

The orbit recovery problem : special cases

- Learning a bag of numbers : $\theta^* \in \mathbb{R}^p, \mathcal{G} = S_p \subset O(p)$
- Learning a rigid body : $\theta^* \in \mathbb{R}^{k \times N}, \mathcal{G} = SO(k)$, acting diagonally on the columns of $\mathbb{R}^{k \times N}$
- Multi Reference Alignment (MRA) : $\theta^* \in \mathbb{R}^p, \mathcal{G} = \mathbb{Z}/p\mathbb{Z}$, acting as cyclic shifts on the coordinates of $\mathbb{R}^p$
- Spherical registration problem : Learn $f : \mathbb{S}^2 \to \mathbb{R}$ from noisy measurements of $f(g^{-1}\bullet)$ with $g \in SO(3)$

Other variants for cryo-EM:

• Additional linear mapping, i.e. $Y_i = \Pi(G_i \cdot \theta^*) + \xi_i$

• Heterogeneity, i.e. we have a finite set $\{\theta^*{}_1, \ldots, \theta^*{}_K\}$, and $Y_i = \Pi(G_i \cdot \theta^*{}_{k(i)}) + \xi_i$ where $k(i) \sim Unif([K])$.

Notions of recovery

The metric

$$d_{\mathcal{G}}(\theta_1, \theta_2) = \min_{g \in \mathcal{G}} \|\theta_1 - g \cdot \theta_2\| = \mathrm{dist}(\theta_1, \mathcal{O}_{\theta_2})$$

Generic signals vs worst case signals

Study the properties of this model for all possible (i.e., worst case) signals vs *generic* signals (i.e., leave out a set of signals of measure zero).

Natural questions

Questions

- Recovery How to perform recovery of $\mathcal{O}_{\theta^*}$ to a given level of accuracy ?
- Sample complexity How many observations n to we need to perform this recovery at a given accuracy level ?
- Optimality How many observations are minimally needed (information theoretic lower bound) ?
- Computational complexity How to perform recovery fast (e.g., in polynomial time in the problem parameters) ? Is there a computational-statistical gap in this model ?

Synchronization

Synchronization is a natural approach to the orbit recovery problem, trying to first "find" the G_i-s (up to trivial symmetries), and then using them to recover $\mathcal{O}_{\theta^*}$.

Synchronization

Synchronization is a natural approach to the orbit recovery problem, trying to first "find" the G_i-s (up to trivial symmetries), and then using them to recover $\mathcal{O}_{\theta^*}$. Concretely, we attempt to find $\{H_i\}_{i=1}^n$ which best *synchronize* the observations $\{Y_i\}_{i=1}^n$, by solving the optimization problem over the group $\mathcal{G}$ given by

$$\min_{H_1,\ldots,H_n\in\mathcal{G}} \sum_{1\le i,j\le n} \|H_i^{-1}Y_i - H_j^{-1}Y_j\|^2.$$

Synchronization is a natural approach to the orbit recovery problem, trying to first "find" the G_i-s (up to trivial symmetries), and then using them to recover $\mathcal{O}_{\theta^*}$. Concretely, we attempt to find $\{H_i\}_{i=1}^n$ which best *synchronize* the observations $\{Y_i\}_{i=1}^n$, by solving the optimization problem over the group $\mathcal{G}$ given by

$$\min_{H_1,\ldots,H_n\in\mathcal{G}} \sum_{1\le i,j\le n} \|H_i^{-1}Y_i - H_j^{-1}Y_j\|^2.$$

Then we approximate $\mathcal{O}_{\theta^*}$ via

$$\hat{\theta} := \frac{1}{n}\sum_{i=1}^{n} \hat{H}_i^{-1}Y_i.$$

Synchronization

Synchronization is a natural approach to the orbit recovery problem, trying to first "find" the G_i-s (up to trivial symmetries), and then using them to recover $\mathcal{O}_{\theta^*}$. Concretely, we attempt to find $\{H_i\}_{i=1}^n$ which best *synchronize* the observations $\{Y_i\}_{i=1}^n$, by solving the optimization problem over the group $\mathcal{G}$ given by

$$\min_{H_1,\dots,H_n \in \mathcal{G}} \sum_{1 \le i,j \le n} \|H_i^{-1} Y_i - H_j^{-1} Y_j\|^2.$$

Then we approximate $\mathcal{O}_{\theta^*}$ via

$$\hat{\theta} := \frac{1}{n} \sum_{i=1}^{n} \hat{H}_i^{-1} Y_i.$$

Problem

!! Synchronization works only in the low noise regime

In the high noise regime, no consistent estimation of the G_i is possible ! [Aguerrebere, Delbracio, Bartesaghi, Sapiro '16].

What can we estimate well ?

Observation

Any function of θ^* that is *invariant* under the action of the group $\mathcal{G}$ can be estimated well using classical statistical methods

What can we estimate well ?

Observation

Any function of θ^* that is *invariant* under the action of the group $\mathcal{G}$ can be estimated well using classical statistical methods

Examples

- For learning a bag of numbers ($\mathcal{G} = S_p$), the classical moments $\mu_k = \sum_{i=1}^{p} \theta_i^k$, for $k \geq 1$

What can we estimate well ?

Observation

Any function of θ^* that is *invariant* under the action of the group $\mathcal{G}$ can be estimated well using classical statistical methods

Examples

- For learning a bag of numbers ($\mathcal{G} = S_p$), the classical moments $\mu_k = \sum_{i=1}^p \theta_i^k$, for $k \geq 1$
- For MRA ($\mathcal{G} = \mathbb{Z}/p\mathbb{Z}$), the classical moments $\sum_{i=1}^p \theta_i^k$, plus additional functions, such as $\sum_{i \in \mathbb{Z}/p\mathbb{Z}} \theta_i \theta_{i+1}$ ...

How far can we reach with invariant functions ?

Enter Invariant Theory

The theory of polynomials that are invariant under the action of a group

Invariant theory

• Let $\mathbf{x} = (x_1, \ldots, x_p)$, and $\mathbb{R}[\mathbf{x}]$ be the ring of polynomials with real coefficients.

• $\mathbb{R}[\mathbf{x}]^{\mathcal{G}}$ denotes the ring of polynomials that are invariant under the action of the group $\mathcal{G}$, via the map $\mathbf{x} \mapsto g.\mathbf{x}$ for $g \in \mathcal{G}$.

- Let $\mathbf{x} = (x_1, \ldots, x_p)$, and $\mathbb{R}[\mathbf{x}]$ be the ring of polynomials with real coefficients.
- $\mathbb{R}[\mathbf{x}]^{\mathcal{G}}$ denotes the ring of polynomials that are invariant under the action of the group $\mathcal{G}$, via the map $\mathbf{x} \mapsto g.\mathbf{x}$ for $g \in \mathcal{G}$.
- Let $U \subseteq \mathbb{R}[\mathbf{x}]^{\mathcal{G}}$ be a subspace of invariant polynomials that we have access to, e.g. can estimate effectively.

Question

Do the values $\{f(\theta^*) : f \in U\}$ determine $\mathcal{O}_{\theta^*}$?

Theorem

The full invariant ring $\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ identifies $\mathcal{O}_\theta$ for every $\theta \in \mathbb{R}^p$.

Theorem

The full invariant ring $\mathbb{R}[\mathbf{x}]^{\mathcal{G}}$ *identifies* $\mathcal{O}_\theta$ *for every* $\theta \in \mathbb{R}^p$.

Definition

The *Reynold's Operator* $\mathcal{R} : \mathbb{R}[\mathbf{x}] \to \mathbb{R}[\mathbf{x}]^{\mathcal{G}}$ is defined by

$$\mathcal{R}(f) := \mathbb{E}_{g \sim \mathrm{Haar}(\mathcal{G})} [g \cdot f] .$$

Invariant theory

Theorem

The full invariant ring $\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ *identifies* $\mathcal{O}_\theta$ *for every* $\theta \in \mathbb{R}^p$.

Theorem

The full invariant ring $\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ identifies $\mathcal{O}_\theta$ for every $\theta \in \mathbb{R}^p$.

Proof.

- Let $\mathfrak{o}_1$ and $\mathfrak{o}_2$ be two distinct (and therefore disjoint) orbits.

Theorem

The full invariant ring $\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ identifies $\mathcal{O}_\theta$ for every $\theta \in \mathbb{R}^p$.

Proof.

- Let $\mathfrak{o}_1$ and $\mathfrak{o}_2$ be two distinct (and therefore disjoint) orbits.
- $\mathfrak{o}_1$ and $\mathfrak{o}_2$ are compact sets, via compactness of $\mathcal{G}$.

Theorem

The full invariant ring $\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ identifies $\mathcal{O}_\theta$ for every $\theta \in \mathbb{R}^p$.

Proof.

- Let $\mathfrak{o}_1$ and $\mathfrak{o}_2$ be two distinct (and therefore disjoint) orbits.
- $\mathfrak{o}_1$ and $\mathfrak{o}_2$ are compact sets, via compactness of $\mathcal{G}$.
- By Urysohn's Lemma, there exists a continuous function $\overline{f} \colon \mathbb{R}^p \to \mathbb{R}$ such that $\overline{f}$ is 0 on $\mathfrak{o}_1$ and 1 on $\mathfrak{o}_2$.

Theorem

The full invariant ring $\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ identifies $\mathcal{O}_\theta$ for every $\theta \in \mathbb{R}^p$.

Proof.

- Let $\mathfrak{o}_1$ and $\mathfrak{o}_2$ be two distinct (and therefore disjoint) orbits.
- $\mathfrak{o}_1$ and $\mathfrak{o}_2$ are compact sets, via compactness of $\mathcal{G}$.
- By Urysohn's Lemma, there exists a continuous function $\bar{f} \colon \mathbb{R}^p \to \mathbb{R}$ such that $\bar{f}$ is 0 on $\mathfrak{o}_1$ and 1 on $\mathfrak{o}_2$.
- By Stone-Weierstrass Theorem, we can approximate $\bar{f}$ to arbitrary accuracy by a polynomial f on any compact subset $K \subset \mathbb{R}^p$ such that $\mathfrak{o}_1 \cup \mathfrak{o}_2 \subseteq K$

Invariant theory

Theorem

The full invariant ring $\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ identifies $\mathcal{O}_\theta$ for every $\theta \in \mathbb{R}^p$.

Proof.

- Let $\mathfrak{o}_1$ and $\mathfrak{o}_2$ be two distinct (and therefore disjoint) orbits.
- $\mathfrak{o}_1$ and $\mathfrak{o}_2$ are compact sets, via compactness of $\mathcal{G}$.
- By Urysohn's Lemma, there exists a continuous function $\overline{f}\colon \mathbb{R}^p \to \mathbb{R}$ such that $\overline{f}$ is 0 on $\mathfrak{o}_1$ and 1 on $\mathfrak{o}_2$.
- By Stone-Weierstrass Theorem, we can approximate $\overline{f}$ to arbitrary accuracy by a polynomial f on any compact subset $K \subset \mathbb{R}^p$ such that $\mathfrak{o}_1 \cup \mathfrak{o}_2 \subseteq K$; let $f \leq 1/3$ on $\mathfrak{o}_1$ and $f \geq 2/3$ on $\mathfrak{o}_2$.

Invariant theory

Theorem

The full invariant ring $\mathbb{R}[\mathbf{x}]^{\mathcal{G}}$ identifies $\mathcal{O}_\theta$ for every $\theta \in \mathbb{R}^p$.

Proof.

- Let $\mathfrak{o}_1$ and $\mathfrak{o}_2$ be two distinct (and therefore disjoint) orbits.
- $\mathfrak{o}_1$ and $\mathfrak{o}_2$ are compact sets, via compactness of $\mathcal{G}$.
- By Urysohn's Lemma, there exists a continuous function $\bar{f}\colon \mathbb{R}^p \to \mathbb{R}$ such that $\bar{f}$ is 0 on $\mathfrak{o}_1$ and 1 on $\mathfrak{o}_2$.
- By Stone-Weierstrass Theorem, we can approximate $\bar{f}$ to arbitrary accuracy by a polynomial f on any compact subset $K \subset \mathbb{R}^p$ such that $\mathfrak{o}_1 \cup \mathfrak{o}_2 \subseteq K$; let $f \leq 1/3$ on $\mathfrak{o}_1$ and $f \geq 2/3$ on $\mathfrak{o}_2$.
- $\mathcal{R}(f)$ is then a $\mathcal{G}$-invariant polynomial which satisfies $\mathcal{R}(f) \leq 1/3$ on $\mathfrak{o}_1$ and $\mathcal{R}(f) \geq 2/3$ on $\mathfrak{o}_2$, thereby separating the orbits $\mathfrak{o}_1$ and $\mathfrak{o}_2$.

$\square$

Transcendence degrees

Algebraic independence

Polynomials $f_1, \ldots, f_m \in \mathbb{R}[\mathbf{x}]$ are algebraically independent if there *does not* exist any non-zero polynomial P in m variables such that $P(f_1, \ldots, f_m) \equiv 0$.

Transcendence degrees

Algebraic independence

Polynomials $f_1, \ldots, f_m \in \mathbb{R}[\mathbf{x}]$ are algebraically independent if there *does not* exist any non-zero polynomial P in m variables such that $P(f_1, \ldots, f_m) \equiv 0$.

Transcendence degree

For a subspace $U \subseteq \mathbb{R}[\mathbf{x}]$, the transcendence degree $\mathrm{trdeg}(U)$ is the maximum possible size of an algebraically independent subset of U.

Transcendence degrees

Algebraic independence

Polynomials $f_1, \ldots, f_m \in \mathbb{R}[\mathbf{x}]$ are algebraically independent if there *does not* exist any non-zero polynomial P in m variables such that $P(f_1, \ldots, f_m) \equiv 0$.

Transcendence degree

For a subspace $U \subseteq \mathbb{R}[\mathbf{x}]$, the transcendence degree $\mathrm{trdeg}(U)$ is the maximum possible size of an algebraically independent subset of U.

- Intuitively, $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$ is the minimal number of parameters required to describe an orbit of $\mathcal{G}$, and is known to be always finite.

Transcendence degrees

Algebraic independence

Polynomials $f_1, \ldots, f_m \in \mathbb{R}[\mathbf{x}]$ are algebraically independent if there *does not* exist any non-zero polynomial P in m variables such that $P(f_1, \ldots, f_m) \equiv 0$.

Transcendence degree

For a subspace $U \subseteq \mathbb{R}[\mathbf{x}]$, the transcendence degree $\mathrm{trdeg}(U)$ is the maximum possible size of an algebraically independent subset of U.

• Intuitively, $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$ is the minimal number of parameters required to describe an orbit of $\mathcal{G}$, and is known to be always finite.
Example : If $\mathcal{G}$ is a finite group, $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}}) = p$.

Generic Recovery

Theorem (Bandeira, Blum-Smith, Kileel, Niles-Weed, Perry, Wein '23)

Let $U \subseteq \mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ be a finite dimensional subspace. If $trdeg(U) = trdeg(\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}})$, then U identifies a generic θ^*.

Generic Recovery

Theorem (Bandeira, Blum-Smith, Kileel, Niles-Weed, Perry, Wein '23)

Let $U \subseteq \mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ be a finite dimensional subspace. If trdeg$(U) =$ trdeg$(\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}})$, then U identifies a generic θ^*. The converse is also true.

Generic Recovery

Theorem (Bandeira, Blum-Smith, Kileel, Niles-Weed, Perry, Wein '23)

Let $U \subseteq \mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ be a finite dimensional subspace. If $trdeg(U) = trdeg(\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}})$, then U identifies a generic θ^*. The converse is also true.

Algorithm to compute transcendence degree

There is an efficient algorithm to compute $\mathrm{trdeg}(U)$ for any finite dimensional subspace $U \subseteq \mathbb{R}[\mathbf{x}]$.

Generic Recovery

Theorem (Bandeira, Blum-Smith, Kileel, Niles-Weed, Perry, Wein '23)

Let $U \subseteq \mathbb{R}[\boldsymbol{x}]^{\mathcal{G}}$ be a finite dimensional subspace. If $trdeg(U) = trdeg(\mathbb{R}[\boldsymbol{x}]^{\mathcal{G}})$, then U identifies a generic θ^*. The converse is also true.

Algorithm to compute transcendence degree

There is an efficient algorithm to compute $\mathrm{trdeg}(U)$ for any finite dimensional subspace $U \subseteq \mathbb{R}[\mathbf{x}]$.

- Based on *rank of Jacobian* criterion for testing algebraic independence
- Based on *matroid structure* of algebraically independent subsets of $\mathbb{R}[\mathbf{x}]$

Moment tensors

Order k moment tensor

The order k moment tensor is defined as

$$T_k(\theta) := \mathbb{E}_{g \sim \mathsf{Haar}(\mathcal{G})}[(g \cdot \theta)^{\otimes k}]$$

Moment tensors

Order k moment tensor

The order k moment tensor is defined as

$$T_k(\theta) := \mathbb{E}_{g\sim\mathsf{Haar}(\mathcal{G})}[(g\cdot\theta)^{\otimes k}]$$

Moment tensors and polynomials

- Each entry of $T_k(\theta)$ is a polynomial in $\mathbb{R}[\mathbf{x}]^{\mathcal{G}}$ that is homogeneous of degree k.
- $T_k(\theta)$ contains the same information as the set of evaluations $\{f(\theta) : f \in \mathbb{R}[\mathbf{x}]^{\mathcal{G}},$ homogeneous of degree $k\}$.

Moment tensors

Order k moment tensor

The order k moment tensor is defined as

$$T_k(\theta) := \mathbb{E}_{g \sim \mathrm{Haar}(\mathcal{G})}[(g \cdot \theta)^{\otimes k}]$$

Moment tensors and polynomials

- Each entry of $T_k(\theta)$ is a polynomial in $\mathbb{R}[\mathbf{x}]^{\mathcal{G}}$ that is homogeneous of degree k.
- $T_k(\theta)$ contains the same information as the set of evaluations $\{f(\theta) : f \in \mathbb{R}[\mathbf{x}]^{\mathcal{G}},\ \text{homogeneous of degree } k\}$.
- In fact, any polynomial in $\mathbb{R}[\mathbf{x}]^{\mathcal{G}}$ that is homogeneous of degree k is a linear combination of the entries of T_k.

Moment tensors and estimation

Estimating $T_k(\theta^*)$

We can estimate $T_k(\theta^*)$ from the given observations by computing

$$\hat{T}_k := \frac{1}{n}\sum_{i=1}^{n}\sum_{g\in G}(g\cdot Y_i)^{\otimes k},$$

correcting for canonical bias terms coming from noise.

Moment tensors and estimation

Estimating $T_k(\theta^*)$

We can estimate $T_k(\theta^*)$ from the given observations by computing

$$\hat{T}_k := \frac{1}{n}\sum_{i=1}^{n}\sum_{g\in G}(g\cdot Y_i)^{\otimes k},$$

correcting for canonical bias terms coming from noise.

Definition

Define $M_{\theta^*,k} := \{\tau \in \mathbb{R}^p : T_i(\tau) = T_i(\theta^*) \forall 1 \leq i \leq k\}$.

Clearly, $\mathcal{O}_{\theta^*} \subseteq M_{\theta^*,k}$.

Moment tensors and estimation

Estimating $T_k(\theta^*)$

We can estimate $T_k(\theta^*)$ from the given observations by computing

$$\hat{T}_k := \frac{1}{n} \sum_{i=1}^{n} \sum_{g \in G} (g \cdot Y_i)^{\otimes k},$$

correcting for canonical bias terms coming from noise.

Definition

Define $M_{\theta^*,k} := \{\tau \in \mathbb{R}^p : T_i(\tau) = T_i(\theta^*) \forall 1 \leq i \leq k\}$.

Clearly, $\mathcal{O}_{\theta^*} \subseteq M_{\theta^*,k}$. For k large enough, $\mathcal{O}_{\theta^*} = M_{\theta^*,k}$.

Moment tensors and estimation

Estimating $T_k(\theta^*)$

We can estimate $T_k(\theta^*)$ from the given observations by computing

$$\hat{T}_k := \frac{1}{n}\sum_{i=1}^{n}\sum_{g\in G}(g\cdot Y_i)^{\otimes k},$$

correcting for canonical bias terms coming from noise.

Definition

Define $M_{\theta^*,k} := \{\tau \in \mathbb{R}^p : T_i(\tau) = T_i(\theta^*) \forall 1 \leq i \leq k\}$.

Clearly, $\mathcal{O}_{\theta^*} \subseteq M_{\theta^*,k}$. For k large enough, $\mathcal{O}_{\theta^*} = M_{\theta^*,k}$.
Alternative estimators via Hermite polynomials.

Moment tensors and estimation

Theorem (Recovering orbits from moments, BBKNPW'23)

We have an explicit estimator $\hat{M}_n(Y_1, \ldots, Y_n)$ (defined via matching empirical moment tensors) such that with high probability it holds that

$$M_{\theta^*,k} \subseteq \hat{M}_n \subseteq M^{\varepsilon}_{\theta^*,k},$$

where $M^{\varepsilon}_{\theta^*,k}$ is the ε-fattening of the set $M_{\theta^*,k}$ for a given tolerance ε and $n = n(\varepsilon)$ observations.

Moment tensors and estimation

Theorem (Recovering orbits from moments, BBKNPW'23)

We have an explicit estimator $\hat{M}_n(Y_1, \dots, Y_n)$ (defined via matching empirical moment tensors) such that with high probability it holds that

$$M_{\theta^*,k} \subseteq \hat{M}_n \subseteq M^{\varepsilon}_{\theta^*,k},$$

where $M^{\varepsilon}_{\theta^*,k}$ is the ε-fattening of the set $M_{\theta^*,k}$ for a given tolerance ε and $n = n(\varepsilon)$ observations.

Sample complexity

$$n = \Omega_{\theta^*,\varepsilon}(\sigma^{2k})$$

Putting everything together : general orbit recovery

A step-by-step procedure

- Compute $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$ (standard techniques depending on $\mathcal{G}$)

Putting everything together : general orbit recovery

A step-by-step procedure

- Compute $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$ (standard techniques depending on $\mathcal{G}$)
- Starting from $j = 1$, consider $U_{\leq j} := \mathrm{Span}(T_1(\mathbf{x}), \ldots, T_j(\mathbf{x}))$

Putting everything together : general orbit recovery

A step-by-step procedure

- Compute $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$ (standard techniques depending on $\mathcal{G}$)
- Starting from $j = 1$, consider $U_{\leq j} := \mathrm{Span}(T_1(\mathbf{x}), \ldots, T_j(\mathbf{x}))$
- Compute $\mathrm{trdeg}(U_{\leq j})$

A step-by-step procedure

- Compute $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$ (standard techniques depending on $\mathcal{G}$)
- Starting from $j = 1$, consider $U_{\leq j} := \mathrm{Span}(T_1(\mathbf{x}), \ldots, T_j(\mathbf{x}))$
- Compute $\mathrm{trdeg}(U_{\leq j})$
- Check if $\mathrm{trdeg}(U_{\leq j}) = \mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$; if yes stop, if no increase j to $j + 1$ and repeat the above steps.

Putting everything together : general orbit recovery

A step-by-step procedure

- Compute $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$ (standard techniques depending on $\mathcal{G}$)
- Starting from $j = 1$, consider $U_{\leq j} := \mathrm{Span}(T_1(\mathbf{x}), \ldots, T_j(\mathbf{x}))$
- Compute $\mathrm{trdeg}(U_{\leq j})$
- Check if $\mathrm{trdeg}(U_{\leq j}) = \mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$; if yes stop, if no increase j to $j + 1$ and repeat the above steps. Let the final index be k, such that $\mathrm{trdeg}(U_{\leq k}) = \mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$.

Putting everything together : general orbit recovery

A step-by-step procedure

- Compute $\text{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$ (standard techniques depending on $\mathcal{G}$)
- Starting from $j = 1$, consider $U_{\leq j} := \text{Span}(T_1(\mathbf{x}), \ldots, T_j(\mathbf{x}))$
- Compute $\text{trdeg}(U_{\leq j})$
- Check if $\text{trdeg}(U_{\leq j}) = \text{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$; if yes stop, if no increase j to $j + 1$ and repeat the above steps. Let the final index be k, such that $\text{trdeg}(U_{\leq k}) = \text{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$.
- For this k, estimate $M_{\theta^*,k}$ (up to accuracy ε) via estimator $\hat{M}_n(Y_1, \ldots, Y_n)$

Putting everything together : general orbit recovery

A step-by-step procedure

- Compute $\text{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$ (standard techniques depending on $\mathcal{G}$)
- Starting from $j = 1$, consider $U_{\leq j} := \text{Span}(T_1(\mathbf{x}), \ldots, T_j(\mathbf{x}))$
- Compute $\text{trdeg}(U_{\leq j})$
- Check if $\text{trdeg}(U_{\leq j}) = \text{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$; if yes stop, if no increase j to $j + 1$ and repeat the above steps. Let the final index be k, such that $\text{trdeg}(U_{\leq k}) = \text{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}})$.
- For this k, estimate $M_{\theta^*,k}$ (up to accuracy ε) via estimator $\hat{M}_n(Y_1, \ldots, Y_n)$
- By the choice of k,the set $M_{\theta^*,k}$ identifies $\mathcal{O}_{\theta^*}$.
- Roughly speaking, invert $\theta \mapsto (T_1(\theta), \ldots, T_k(\theta))$ based on data.

Multi Reference Alignment (MRA)

- $\mathcal{G} = \mathbb{Z}/p\mathbb{Z}$

Multi Reference Alignment (MRA)

- $\mathcal{G} = \mathbb{Z}/p\mathbb{Z}$
-
  - $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}}) = p$
  - $T_1(x)$ has 1 distinct entry
  - $T_2(x)$ has $\lfloor p/2 \rfloor + 1$ distinct entries
  - $T_3(x))$ has $p + \lceil (p-1)(p-2)/6 \rceil$ distinct entries
- Recovery possible for generic signals from 3-rd order moment tensors

Multi Reference Alignment (MRA)

- $\mathcal{G} = \mathbb{Z}/p\mathbb{Z}$
-
  - $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}}) = p$
  - $T_1(x)$ has 1 distinct entry
  - $T_2(x)$ has $\lfloor p/2 \rfloor + 1$ distinct entries
  - $T_3(x))$ has $p + \lceil (p-1)(p-2)/6 \rceil$ distinct entries
- Recovery possible for generic signals from 3-rd order moment tensors
- Sample complexity $O(\sigma^6)$

Multi Reference Alignment (MRA)

- $\mathcal{G} = \mathbb{Z}/p\mathbb{Z}$
-
  - $\text{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}}) = p$
  - $T_1(x)$ has 1 distinct entry
  - $T_2(x)$ has $\lfloor p/2 \rfloor + 1$ distinct entries
  - $T_3(x))$ has $p + \lceil (p-1)(p-2)/6 \rceil$ distinct entries
- Recovery possible for generic signals from 3-rd order moment tensors
- Sample complexity $O(\sigma^6)$
- But most significant regime : $\sigma \uparrow \infty$!

Multi Reference Alignment (MRA)

- $\mathcal{G} = \mathbb{Z}/p\mathbb{Z}$
-
  - $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}}) = p$
  - $T_1(x)$ has 1 distinct entry
  - $T_2(x)$ has $\lfloor p/2 \rfloor + 1$ distinct entries
  - $T_3(x))$ has $p + \lceil (p-1)(p-2)/6 \rceil$ distinct entries
- Recovery possible for generic signals from 3-rd order moment tensors
- Sample complexity $O(\sigma^6)$
- But most significant regime : $\sigma \uparrow \infty$! Need to improve on sample complexity in important structural settings for the signal

Multi Reference Alignment (MRA)

- $\mathcal{G} = \mathbb{Z}/p\mathbb{Z}$
-
    - $\mathrm{trdeg}(\mathbb{R}[\mathbf{x}]^{\mathcal{G}}) = p$
    - $T_1(x)$ has 1 distinct entry
    - $T_2(x)$ has $\lfloor p/2 \rfloor + 1$ distinct entries
    - $T_3(x))$ has $p + \lceil (p-1)(p-2)/6 \rceil$ distinct entries
- Recovery possible for generic signals from 3-rd order moment tensors
- Sample complexity $O(\sigma^6)$
- But most significant regime : $\sigma \uparrow \infty$! Need to improve on sample complexity in important structural settings for the signal

Sample complexity of Sparse Multi Reference Alignment (MRA)

- Sparsity is the most fundamental structural feature for real-world signals
- Fundamental question : How does the sample complexity of sparse MRA scale with σ ?

Sample complexity of Sparse Multi Reference Alignment (MRA)

- Sparsity is the most fundamental structural feature for real-world signals
- Fundamental question : How does the sample complexity of sparse MRA scale with σ ?
- Without latent symmetries, the sample complexity is $O(\sigma^2)$
- Without sparsity, the sample complexity is $O(\sigma^6)$

Sample complexity of Sparse Multi Reference Alignment (MRA)

Theorem (G.-Rigollet,'23)

*The sample complexity of MRA for the MLE exhibits a **novel intermediate scaling** of $O(\sigma^4)$ for generic sparse signals.*

Sample complexity of Sparse Multi Reference Alignment (MRA)

Theorem (G.-Rigollet,'23)

The sample complexity of MRA for the MLE exhibits a ***novel intermediate scaling*** *of $O(\sigma^4)$ for generic sparse signals.*

- $O(\sigma^4)$ scaling is the best possible for generic sparse signals. (G.-Rigollet,'23)

Sample complexity of Sparse Multi Reference Alignment (MRA)

Theorem (G.-Rigollet,'23)

The sample complexity of MRA for the MLE exhibits a ***novel intermediate scaling*** *of $O(\sigma^4)$ for generic sparse signals.*

- $O(\sigma^4)$ scaling is the best possible for generic sparse signals. (G.-Rigollet,'23)
- Without sparsity, $O(\sigma^6)$ is best possible for generic signals. (G.-Rigollet,'23)

Sample complexity of Sparse Multi Reference Alignment (MRA)

Theorem (G.-Rigollet,'23)

The sample complexity of MRA for the MLE exhibits a ***novel intermediate scaling*** *of $O(\sigma^4)$ for generic sparse signals.*

- $O(\sigma^4)$ scaling is the best possible for generic sparse signals. (G.-Rigollet,'23)
- Without sparsity, $O(\sigma^6)$ is best possible for generic signals. (G.-Rigollet,'23)
- Explicit dependence on sparsity level and p. (G.-Rigollet,'23)

Sample complexity of Sparse Multi Reference Alignment (MRA)

Theorem (G.-Rigollet,'23)

The sample complexity of MRA for the MLE exhibits a ***novel intermediate scaling*** *of $O(\sigma^4)$ for generic sparse signals.*

- $O(\sigma^4)$ scaling is the best possible for generic sparse signals. (G.-Rigollet,'23)
- Without sparsity, $O(\sigma^6)$ is best possible for generic signals. (G.-Rigollet,'23)
- Explicit dependence on sparsity level and p. (G.-Rigollet,'23)

Sample complexity of Sparse Multi Reference Alignment (MRA)

Theorem (G.-Tran,'24+)

If sparsity is in Fourier space, then sample complexity is $O(\sigma^6)$ for generic sparse signals

Sample complexity of Sparse Multi Reference Alignment (MRA)

Theorem (G.-Tran,'24+)

If sparsity is in Fourier space, then sample complexity is $O(\sigma^6)$ *for generic sparse signals*

Theorem (G.-Mukherjee-Pan,'24+)

Minimax optimal rates of estimation for sparse MRA in dilute regime of sparsity

Sample complexity of Sparse Multi Reference Alignment (MRA)

- The restricted MLE $\hat{\theta}_{\text{MLE}}$ satisfies a central limit theorem with convergence of $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^*)$ to $N(0, \mathcal{I}(\theta^*)^{-1})$, where $\mathcal{I}(\theta^*)$ is the Fisher information matrix for the model at the true parameter value θ^*.

Sample complexity of Sparse Multi Reference Alignment (MRA)

- The restricted MLE $\hat{\theta}_{\text{MLE}}$ satisfies a central limit theorem with convergence of $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^*)$ to $N(0, \mathcal{I}(\theta^*)^{-1})$, where $\mathcal{I}(\theta^*)$ is the Fisher information matrix for the model at the true parameter value θ^*.
- Thus, $(\hat{\theta}_{\text{MLE}} - \theta^*) \simeq \frac{1}{\sqrt{n}} \cdot \mathcal{I}(\theta^*)^{-1}$

Sample complexity of Sparse Multi Reference Alignment (MRA)

- The restricted MLE $\hat{\theta}_{\text{MLE}}$ satisfies a central limit theorem with convergence of $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^*)$ to $N(0, \mathcal{I}(\theta^*)^{-1})$, where $\mathcal{I}(\theta^*)$ is the Fisher information matrix for the model at the true parameter value θ^*.
- Thus, $(\hat{\theta}_{\text{MLE}} - \theta^*) \simeq \frac{1}{\sqrt{n}} \cdot \mathcal{I}(\theta^*)^{-1} = \frac{1}{\sqrt{n}} \cdot \nabla^2_\theta (D_{KL}(\theta \parallel \theta^*))^{-1}$.

Sample complexity of Sparse Multi Reference Alignment (MRA)

- The restricted MLE $\hat{\theta}_{\text{MLE}}$ satisfies a central limit theorem with convergence of $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^*)$ to $N(0, \mathcal{I}(\theta^*)^{-1})$, where $\mathcal{I}(\theta^*)$ is the Fisher information matrix for the model at the true parameter value θ^*.
- Thus, $(\hat{\theta}_{\text{MLE}} - \theta^*) \simeq \frac{1}{\sqrt{n}} \cdot \mathcal{I}(\theta^*)^{-1} = \frac{1}{\sqrt{n}} \cdot \nabla^2_\theta (D_{KL}(\theta \parallel \theta^*))^{-1}$.
- If the *second moment tensor* mapping $\theta \mapsto T_2(\theta) = \mathbb{E}_{g \sim \text{Haar}(\mathbb{Z}_p)}[(g \cdot (\theta)^{\otimes k}]$ is *suitably non-degenerate* at $\theta = \theta^*$, then $(D_{KL}(\theta \parallel \theta^*))^{-1}$ is $O(\sigma^2)$,

Sample complexity of Sparse Multi Reference Alignment (MRA)

- The restricted MLE $\hat{\theta}_{\mathrm{MLE}}$ satisfies a central limit theorem with convergence of $\sqrt{n}(\hat{\theta}_{\mathrm{MLE}} - \theta^*)$ to $N(0, \mathcal{I}(\theta^*)^{-1})$, where $\mathcal{I}(\theta^*)$ is the Fisher information matrix for the model at the true parameter value θ^*.
- Thus, $(\hat{\theta}_{\mathrm{MLE}} - \theta^*) \simeq \frac{1}{\sqrt{n}} \cdot \mathcal{I}(\theta^*)^{-1} = \frac{1}{\sqrt{n}} \cdot \nabla_\theta^2 (D_{KL}(\theta \parallel \theta^*))^{-1}$.
- If the *second moment tensor* mapping $\theta \mapsto T_2(\theta) = \mathbb{E}_{g \sim \mathrm{Haar}(\mathbb{Z}_p)}[(g \cdot (\theta)^{\otimes k}]$ is *suitably non-degenerate* at $\theta = \theta^*$, then $(D_{KL}(\theta \parallel \theta^*))^{-1}$ is $O(\sigma^2)$, indicating sample complexity $n \sim \sigma^4$.

Sample complexity of Sparse Multi Reference Alignment (MRA)

- Entries of the matrix $T_2(\theta)$ are the *auto-correlations* of the signal θ

Sample complexity of Sparse Multi Reference Alignment (MRA)

- Entries of the matrix $T_2(\theta)$ are the *auto-correlations* of the signal θ
- Non-degeneracy of $\theta \mapsto T_2(\theta) \longleftrightarrow$ Recovery of signal θ from its autocorrelations $\longleftrightarrow$ Recovery of $\hat{\theta}$ from $|\hat{\theta}|$

Sample complexity of Sparse Multi Reference Alignment (MRA)

- Entries of the matrix $T_2(\theta)$ are the *auto-correlations* of the signal θ
- Non-degeneracy of $\theta \mapsto T_2(\theta) \longleftrightarrow$ Recovery of signal θ from its autocorrelations $\longleftrightarrow$ Recovery of $\hat{\theta}$ from $|\hat{\theta}|$
- Crystallographic phase retrieval

Sample complexity of Sparse Multi Reference Alignment (MRA)

- Entries of the matrix $T_2(\theta)$ are the *auto-correlations* of the signal θ
- Non-degeneracy of $\theta \mapsto T_2(\theta) \longleftrightarrow$ Recovery of signal θ from its autocorrelations $\longleftrightarrow$ Recovery of $\hat{\theta}$ from $|\hat{\theta}|$
- Crystallographic phase retrieval
- Support recovery from auto-correlations $\longleftrightarrow$ Beltway problem / Turnpike problem / Partial digest problem

Sample complexity of Sparse Multi Reference Alignment (MRA)

- Entries of the matrix $T_2(\theta)$ are the *auto-correlations* of the signal θ
- Non-degeneracy of $\theta \mapsto T_2(\theta) \longleftrightarrow$ Recovery of signal θ from its autocorrelations $\longleftrightarrow$ Recovery of $\hat{\theta}$ from $|\hat{\theta}|$
- Crystallographic phase retrieval
- Support recovery from auto-correlations $\longleftrightarrow$ Beltway problem / Turnpike problem / Partial digest problem
- Non-degeneracy of $\theta \mapsto T_2(\theta)$ is best analysed in the Fourier space;

Sample complexity of Sparse Multi Reference Alignment (MRA)

- Entries of the matrix $T_2(\theta)$ are the *auto-correlations* of the signal θ
- Non-degeneracy of $\theta \mapsto T_2(\theta) \longleftrightarrow$ Recovery of signal θ from its autocorrelations $\longleftrightarrow$ Recovery of $\hat{\theta}$ from $|\hat{\theta}|$
- Crystallographic phase retrieval
- Support recovery from auto-correlations $\longleftrightarrow$ Beltway problem / Turnpike problem / Partial digest problem
- Non-degeneracy of $\theta \mapsto T_2(\theta)$ is best analysed in the Fourier space; Uniform Uncertainty Principles allow us to switch between physical and Fourier space efficiently , entailing a sparse approximation in the frequency variables.

Sample complexity of Sparse Multi Reference Alignment (MRA)

- Entries of the matrix $T_2(\theta)$ are the *auto-correlations* of the signal θ
- Non-degeneracy of $\theta \mapsto T_2(\theta) \longleftrightarrow$ Recovery of signal θ from its autocorrelations $\longleftrightarrow$ Recovery of $\hat{\theta}$ from $|\hat{\theta}|$
- Crystallographic phase retrieval
- Support recovery from auto-correlations $\longleftrightarrow$ Beltway problem / Turnpike problem / Partial digest problem
- Non-degeneracy of $\theta \mapsto T_2(\theta)$ is best analysed in the Fourier space; Uniform Uncertainty Principles allow us to switch between physical and Fourier space efficiently , entailing a sparse approximation in the frequency variables.

Information geometry

The likelihood of the group invariant learning problem is given by

$$p_\theta(y) = \frac{1}{|\mathcal{G}|} \sum_{R \in \mathcal{G}} \frac{1}{(\sqrt{2\pi}\sigma)^L} \exp\left(-\frac{\|y - R\theta\|_2^2}{2\sigma^2}\right)$$

Information geometry

The likelihood of the group invariant learning problem is given by

$$p_\theta(y) = \frac{1}{|\mathcal{G}|} \sum_{R \in \mathcal{G}} \frac{1}{(\sqrt{2\pi}\sigma)^L} \exp\left(-\frac{\|y - R\theta\|_2^2}{2\sigma^2}\right)$$

The log likelihood corresponding to the data $\{y_1, \ldots, y_n\}$ as

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log p_\theta(y_i).$$

Information geometry

The likelihood of the group invariant learning problem is given by

$$p_\theta(y) = \frac{1}{|\mathcal{G}|} \sum_{R \in \mathcal{G}} \frac{1}{(\sqrt{2\pi}\sigma)^L} \exp\left(-\frac{\|y - R\theta\|_2^2}{2\sigma^2}\right)$$

The log likelihood corresponding to the data $\{y_1, \ldots, y_n\}$ as

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} \log p_\theta(y_i).$$

The population risk of the model is given by

$$R(\theta) = -\mathbb{E}_{p_{\theta_0}}[\log p_\theta(Y)] + C,$$

Information geometry

$$
\begin{aligned}
R(\theta) &= -\int \log p_\theta(y) p_{\theta_0}(y) \mathrm{d}y + C \\
&= \int \log\left(\frac{p_{\theta_0}(y)}{p_\theta(y)} \cdot \frac{1}{p_{\theta_0}(y)}\right) p_{\theta_0}(y) \mathrm{d}y + C \\
&= D_{KL}(p_{\theta_0} || p_\theta) - \left(\int p_{\theta_0}(y) \log p_{\theta_0}(y) \mathrm{d}y\right) + C
\end{aligned}
$$

where $D_{KL}(p_{\theta_0}||p_\theta)$ is the Kullback-Leibler divergence between p_{θ_0} and p_θ.

Information geometry

$$
\begin{aligned}
R(\theta) &= -\int \log p_\theta(y) p_{\theta_0}(y) \mathrm{d}y + C \\
&= \int \log \left(\frac{p_{\theta_0}(y)}{p_\theta(y)} \cdot \frac{1}{p_{\theta_0}(y)} \right) p_{\theta_0}(y) \mathrm{d}y + C \\
&= D_{KL}(p_{\theta_0} || p_\theta) - \left(\int p_{\theta_0}(y) \log p_{\theta_0}(y) \mathrm{d}y \right) + C
\end{aligned}
$$

where $D_{KL}(p_{\theta_0}||p_\theta)$ is the Kullback-Leibler divergence between p_{θ_0} and p_θ. Since θ_0 is fixed, as a function of θ, the population risk $R(\theta)$ equals

$$
R(\theta) = D_{KL}(p_{\theta_0}||p_\theta) + C(\theta_0),
$$

where $C(\theta_0)$ is a function of θ_0.

Information geometry

The Fisher information matrix of the MRA model is given by

$$I(\theta_0) = -\mathbb{E}[\nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta_0}] = \nabla_\theta^2 R(\theta_0),$$

where ∇_θ^2 denotes the Hessian with respect to the variable θ.

The Fisher information matrix of the MRA model is given by

$$I(\theta_0) = -\mathbb{E}[\nabla_\theta^2 \log p_\theta(Y)\big|_{\theta=\theta_0}] = \nabla_\theta^2 R(\theta_0),$$

where ∇_θ^2 denotes the Hessian with respect to the variable θ.

Theorem (Abbe,Bendory,Leeb,Pereira,Sharon,Singer'18)

The MLE $\tilde{\theta}_n$ is an asymptotically consistent estimate for the true signal θ_0 in the MRA model.

Information geometry

This immediately enables us to invoke standard asymptotic normality theory for MLEs (c.f. van der Vaart):

Theorem

$\sqrt{n}(\tilde{\theta} - \theta_0)$ *is asymptotically normal with and covariance* $I(\theta_0)^{-1}$.

This immediately enables us to invoke standard asymptotic normality theory for MLEs (c.f. van der Vaart):

Theorem

$\sqrt{n}(\tilde{\theta} - \theta_0)$ is asymptotically normal with and covariance $I(\theta_0)^{-1}$.

Upshot: The distance $\rho(\tilde{\theta}_n, \theta_0)$ is of the order

$$n^{-1/2}\sqrt{\mathrm{Tr}\left[I(\theta)^{-1}\right]} = n^{-1/2}\sqrt{\mathrm{Tr}\left[\left[\nabla^2_{\theta|\theta=\theta_0} D_{KL}(p_{\theta_0}||p_\theta)\right]^{-1}\right]}.$$

Theorem (Bandeira, Niles-Weed, Rigollet'20)

Let $\theta, \varphi \in \mathbb{R}^p$ *satisfy* $3\rho(\theta, \varphi) \leq \|\theta\| \leq \sigma$ *and* $\mathbb{E}_{\mathcal{G}}[G\theta] = \mathbb{E}_{\mathcal{G}}[G\varphi] = 0$.
Let $\Delta_m = \Delta_m(\theta, \varphi) = \mathbb{E}[(G\theta)^{\otimes m}] - \mathbb{E}[(G\varphi)^{\otimes m}]$.

Theorem (Bandeira,Niles-Weed,Rigollet'20)

Let $\theta, \varphi \in \mathbb{R}^p$ *satisfy* $3\rho(\theta, \varphi) \le \|\theta\| \le \sigma$ *and* $\mathbb{E}_{\mathcal{G}}[G\theta] = \mathbb{E}_{\mathcal{G}}[G\varphi] = 0$.
Let $\Delta_m = \Delta_m(\theta, \varphi) = \mathbb{E}[(G\theta)^{\otimes m}] - \mathbb{E}[(G\varphi)^{\otimes m}]$.
For any $k \ge 1$, *there exist universal constants* $\underline{C}$ *and* $\overline{C}$ *such that*

$$\underline{C} \sum_{m=1}^{\infty} \frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m} m!} \le D_{KL}(p_\theta || p_\varphi)$$

and

$$D_{KL}(p_\theta || p_\varphi) \le 2 \sum_{m=1}^{k-1} \frac{\|\Delta_m\|^2}{\sigma^{2m} m!} + \overline{C} \frac{\|\theta\|^{2k-2} \rho(\theta, \varphi)^2}{\sigma^{2k}}.$$

Corollary

If j is the minimum index such that $\|\Delta_j(\theta, \theta_0)\| \gtrsim \rho(\theta, \theta_0)$ on a neighbourhood of θ_0, then sample complexity scales as σ^{2j}.

Corollary

If j is the minimum index such that $\|\Delta_j(\theta, \theta_0)\| \gtrsim \rho(\theta, \theta_0)$ on a neighbourhood of θ_0, then sample complexity scales as σ^{2j}. Upshot: to improve sample complexity beyond σ^6, need to show non-degeneracy of $\theta \mapsto \|\Delta_j(\theta, \theta_0)\|$ on a neighbourhood of σ.

The Bernoulli Gaussian model

Definition (Generic sparse signals)

Generic support : Independent Bernoulli (s/p) sampling

The Bernoulli Gaussian model

Definition (Generic sparse signals)

Generic support : Independent Bernoulli (s/p) sampling
Generic values : Independent Gaussians

The beltway problem

Definition

A subset $S \subseteq \mathbb{Z}$ is said to be collision-free if its pairwise differences $D := \{i - j : i, j \in D\}$ are unique.

The beltway problem

Definition

A subset $S \subseteq \mathbb{Z}$ is said to be collision-free if its pairwise differences $D := \{i - j : i, j \in D\}$ are unique.

Question (Beltway Problem / Turnpike Problem / Partial Digest Problem (computational biology, signal processing))

What can we say about the set S from its pairwise differences D?

The beltway problem

Definition

A subset $S \subseteq \mathbb{Z}$ is said to be collision-free if its pairwise differences $D := \{i - j : i, j \in D\}$ are unique.

Question (Beltway Problem / Turnpike Problem / Partial Digest Problem (computational biology, signal processing))

What can we say about the set S from its pairwise differences D?

Conjecture (Piccard'39)

If S is collision free, D determines S uniquely up to trivial symmetries.

The beltway problem

Definition

A subset $S \subseteq \mathbb{Z}$ is said to be collision-free if its pairwise differences $D := \{i - j : i, j \in D\}$ are unique.

Question (Beltway Problem / Turnpike Problem / Partial Digest Problem (computational biology, signal processing))

What can we say about the set S from its pairwise differences D?

Conjecture (Piccard'39)

If S is collision free, D determines S uniquely up to trivial symmetries.

Theorem (Bekir,Golomb'04'07;Bloom'77)

Piccard's conjecture is true for $|S| \geq 7$.

The *dilute* regime of sparsity

- For $s = o(p^{1/4})$, a generic support is collision-free with high probability

The *dilute* regime of sparsity

- For $s = o(p^{1/4})$, a generic support is collision-free with high probability
- For small h, we have $\Delta(\theta_0 + h, \theta_0) = \mathbb{E}_{\mathcal{G}}[G\theta_0 h^* G^* + Gh\theta_0^* G^*] =: J$, to the leading order

The *dilute* regime of sparsity

- For $s = o(p^{1/4})$, a generic support is collision-free with high probability
- For small h, we have $\Delta(\theta_0 + h, \theta_0) = \mathbb{E}_{\mathcal{G}}[G\theta_0 h^* G^* + Gh\theta_0^* G^*] =: J$, to the leading order
- (i,j) entry of J is $\frac{1}{p}\sum_{g=1}^{p}[\theta_0(i+g)h(j+g) + h(i+g)\theta_0(j+g)]$

The *dilute* regime of sparsity

- For $s = o(p^{1/4})$, a generic support is collision-free with high probability
- For small h, we have $\Delta(\theta_0 + h, \theta_0) = \mathbb{E}_{\mathcal{G}}[G\theta_0 h^* G^* + Gh\theta_0^* G^*] =: J$, to the leading order
- (i, j) entry of J is $\frac{1}{p}\sum_{g=1}^{p}[\theta_0(i+g)h(j+g) + h(i+g)\theta_0(j+g)]$
- J is Toeplitz, i.e. $J_{ij} = J_{i-j}$

The *dilute* regime of sparsity

- For $s = o(p^{1/4})$, a generic support is collision-free with high probability
- For small h, we have $\Delta(\theta_0 + h, \theta_0) = \mathbb{E}_{\mathcal{G}}[G\theta_0 h^* G^* + Gh\theta_0^* G^*] =: J$, to the leading order
- (i,j) entry of J is $\frac{1}{p}\sum_{g=1}^{p}[\theta_0(i+g)h(j+g) + h(i+g)\theta_0(j+g)]$
- J is Toeplitz, i.e. $J_{ij} = J_{i-j}$
- Target signal not too small on its support

The *dilute* regime of sparsity

- For $s = o(p^{1/4})$, a generic support is collision-free with high probability
- For small h, we have $\Delta(\theta_0 + h, \theta_0) = \mathbb{E}_{\mathcal{G}}[G\theta_0 h^* G^* + G h \theta_0^* G^*] =: J$, to the leading order
- (i, j) entry of J is $\frac{1}{p} \sum_{g=1}^{p} [\theta_0(i+g) h(j+g) + h(i+g)\theta_0(j+g)]$
- J is Toeplitz, i.e. $J_{ij} = J_{i-j}$
- Target signal not too small on its support $\implies$ θ_0, h have same support S

The *dilute* regime of sparsity

- For $s = o(p^{1/4})$, a generic support is collision-free with high probability
- For small h, we have $\Delta(\theta_0 + h, \theta_0) = \mathbb{E}_{\mathcal{G}}[G\theta_0 h^* G^* + Gh\theta_0^* G^*] =: J$, to the leading order
- (i,j) entry of J is $\frac{1}{p}\sum_{g=1}^{p}[\theta_0(i+g)h(j+g) + h(i+g)\theta_0(j+g)]$
- J is Toeplitz, i.e. $J_{ij} = J_{i-j}$
- Target signal not too small on its support $\implies$ θ_0, h have same support S
- $J_{ij} = 0$ unless both i,j belong to support S ($\iff$ $i - j \in D$)

The *dilute* regime of sparsity

- For $s = o(p^{1/4})$, a generic support is collision-free with high probability
- For small h, we have $\Delta(\theta_0 + h, \theta_0) = \mathbb{E}_{\mathcal{G}}[G\theta_0 h^* G^* + Gh\theta_0^* G^*] =: J$, to the leading order
- (i,j) entry of J is $\frac{1}{p}\sum_{g=1}^{p}[\theta_0(i+g)h(j+g) + h(i+g)\theta_0(j+g)]$
- J is Toeplitz, i.e. $J_{ij} = J_{i-j}$
- Target signal not too small on its support $\implies$ θ_0, h have same support S
- $J_{ij} = 0$ unless both i,j belong to support S ($\iff$ $i - j \in D$)
- S collision-free $\implies$ exactly one term in $\sum_{g=1}^{p}[\theta_0(i+g)h(j+g) + h(i+g)\theta_0(j+g)]$ is non-zero $\implies$ linear lower bound in h.

The *moderate* regime of sparsity

- $\mathrm{polylog}(p) \lesssim s \lesssim p/\mathrm{polylog}(p)$
- Signal θ_0 is symmetric (implies Fourier coefficients are real)

The *moderate* regime of sparsity

- $\mathrm{polylog}(p) \lesssim s \lesssim p/\mathrm{polylog}(p)$
- Signal θ_0 is symmetric (implies Fourier coefficients are real)
- Set $\check{h}(x) = h(-x)$, then

$$\frac{1}{p}\sum_{g=1}^{p} \theta_0(i+g)h(j+g) = \frac{1}{p}\sum_{g=1}^{p} \theta_0(i+g)\check{h}(-j-g) = [\theta_0 * \check{h}](i-j).$$

The *moderate* regime of sparsity

- $\mathrm{polylog}(p) \lesssim s \lesssim p/\mathrm{polylog}(p)$
- Signal θ_0 is symmetric (implies Fourier coefficients are real)
- Set $\check{h}(x) = h(-x)$, then

$$\frac{1}{p}\sum_{g=1}^{p} \theta_0(i+g)h(j+g) = \frac{1}{p}\sum_{g=1}^{p} \theta_0(i+g)\check{h}(-j-g) = [\theta_0 * \check{h}](i-j).$$

- Set $\mathcal{M}[v] := \big(v(i-j)\big)$, then

$$\Delta(\theta_0+h, \theta_0) = \mathbb{E}_{\mathcal{G}}[G\theta_0 h^* G^* + Gh\theta_0^* G^*] + o(h) = \mathcal{M}[\theta_0 * \check{h}] + \mathcal{M}[\check{\theta}_0 * h] + o(h)$$

The *moderate* regime of sparsity

- $\text{polylog}(p) \lesssim s \lesssim p/\text{polylog}(p)$
- Signal θ_0 is symmetric (implies Fourier coefficients are real)
- Set $\check{h}(x) = h(-x)$, then

$$\frac{1}{p}\sum_{g=1}^{p}\theta_0(i+g)h(j+g) = \frac{1}{p}\sum_{g=1}^{p}\theta_0(i+g)\check{h}(-j-g) = [\theta_0 * \check{h}](i-j).$$

- Set $\mathcal{M}[v] := \big(v(i-j)\big)$, then

$$\Delta(\theta_0+h,\theta_0) = \mathbb{E}_{\mathcal{G}}[G\theta_0 h^* G^* + Gh\theta_0^* G^*] + o(h) = \mathcal{M}[\theta_0 * \check{h}] + \mathcal{M}[\check{\theta}_0 * h] + o(h)$$

- Discrete Fourier analysis and Parseval's Theorem:

$$\|\mathcal{M}(\theta_0 * \check{h})\|_F = \sqrt{p}\|\theta_0 * \check{h}\|_2 = \sqrt{p}\cdot\frac{1}{\sqrt{p}}\cdot\|\widehat{\theta_0 * \check{h}}\|_2 = \|\hat{\theta}_0 \cdot \hat{\check{h}}\|_2 = \|\hat{\theta}_0 \cdot \overline{\hat{h}}\|_2$$

- All said and done :
  $\|\Delta(\theta_0 + h, \theta_0)\|^2 = \sum_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|^2 |\hat{h}(\xi)|^2$

The *moderate* regime of sparsity

- All said and done :
  $\|\Delta(\theta_0 + h, \theta_0)\|^2 = \sum_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|^2 |\hat{h}(\xi)|^2$
- Naive bound : lower bound $\min_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|$ ... too crude

- All said and done :
  $\|\Delta(\theta_0 + h, \theta_0)\|^2 = \sum_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|^2 |\hat{h}(\xi)|^2$
- Naive bound : lower bound $\min_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|$ ... too crude
- Want to leverage sparsity

The *moderate* regime of sparsity

- All said and done : $\|\Delta(\theta_0 + h, \theta_0)\|^2 = \sum_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|^2 |\hat{h}(\xi)|^2$
- Naive bound : lower bound $\min_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|$ ... too crude
- Want to leverage sparsity which is in physical coordinates

The *moderate* regime of sparsity

- All said and done : $\|\Delta(\theta_0 + h, \theta_0)\|^2 = \sum_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|^2 |\hat{h}(\xi)|^2$
- Naive bound : lower bound $\min_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|$ ... too crude
- Want to leverage sparsity which is in physical coordinates but analysis is in Fourier coordinates

The *moderate* regime of sparsity

- All said and done :
  $\|\Delta(\theta_0 + h, \theta_0)\|^2 = \sum_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|^2 |\hat{h}(\xi)|^2$
- Naive bound : lower bound $\min_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|$ ... too crude
- Want to leverage sparsity which is in physical coordinates but analysis is in Fourier coordinates
- Need: a bridge between physical and Fourier coordinates that

The *moderate* regime of sparsity

- All said and done :
  $\|\Delta(\theta_0 + h, \theta_0)\|^2 = \sum_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|^2 |\hat{h}(\xi)|^2$
- Naive bound : lower bound $\min_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|$ ... too crude
- Want to leverage sparsity which is in physical coordinates but analysis is in Fourier coordinates
- Need: a bridge between physical and Fourier coordinates that
  - (a) doesn't lose much information

The *moderate* regime of sparsity

- All said and done : $\|\Delta(\theta_0 + h, \theta_0)\|^2 = \sum_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|^2 |\hat{h}(\xi)|^2$
- Naive bound : lower bound $\min_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|$ ... too crude
- Want to leverage sparsity which is in physical coordinates but analysis is in Fourier coordinates
- Need: a bridge between physical and Fourier coordinates that
  - (a) doesn't lose much information
  - (b) transfers sparsity to Fourier coordinates (e..g, so that $\min_{\xi \in \Lambda} |\hat{\theta}_0(\xi)|$ is not too small)

The *moderate* regime of sparsity

- All said and done :
  $\|\Delta(\theta_0 + h, \theta_0)\|^2 = \sum_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|^2 |\hat{h}(\xi)|^2$
- Naive bound : lower bound $\min_{\xi \in \mathbb{Z}/p\mathbb{Z}} |\hat{\theta}_0(\xi)|$ ... too crude
- Want to leverage sparsity which is in physical coordinates but analysis is in Fourier coordinates
- Need: a bridge between physical and Fourier coordinates that
  - (a) doesn't lose much information
  - (b) transfers sparsity to Fourier coordinates (e..g, so that $\min_{\xi \in \Lambda} |\hat{\theta}_0(\xi)|$ is not too small)

The *moderate* regime of sparsity

- Need: a bridge between physical and Fourier coordinates that
  - (a) doesn't lose much information
  - (b) transfers sparsity to Fourier coordinates (e..g, so that $\min_{\xi \in \Lambda} |\hat{\theta}_0(\xi)|$ is not too small)

The *moderate* regime of sparsity

- Need: a bridge between physical and Fourier coordinates that
  - (a) doesn't lose much information
  - (b) transfers sparsity to Fourier coordinates (e..g, so that $\min_{\xi \in \Lambda} |\hat{\theta}_0(\xi)|$ is not too small)
- Solution: Uniform Uncertainty Principle (UUP) : random set of frequencies Λ of size $s \log p$ suffices for (a) with high probability

The *moderate* regime of sparsity

- Need: a bridge between physical and Fourier coordinates that
  - (a) doesn't lose much information
  - (b) transfers sparsity to Fourier coordinates (e..g, so that $\min_{\xi \in \Lambda} |\hat{\theta}_0(\xi)|$ is not too small)
- Solution: Uniform Uncertainty Principle (UUP) : random set of frequencies Λ of size $s \log p$ suffices for (a) with high probability
- But for (b), min of $\hat{\theta}_0$ over a random set of frequencies Λ is still very small with high probability (in Λ

The *moderate* regime of sparsity

- Need: a bridge between physical and Fourier coordinates that
  - (a) doesn't lose much information
  - (b) transfers sparsity to Fourier coordinates (e..g, so that $\min_{\xi \in \Lambda} |\hat{\theta}_0(\xi)|$ is not too small)
- Solution: Uniform Uncertainty Principle (UUP) : random set of frequencies Λ of size $s \log p$ suffices for (a) with high probability
- But for (b), min of $\hat{\theta}_0$ over a random set of frequencies Λ is still very small with high probability (in Λ
- Show that this high probability is strictly smaller than 1
- Application of probabilistic method to show existence of good set Λ of frequencies satisfying both (a) and (b) where the *probability of finding good set* $\to 0$ *with system size*

Information geometry : the upper bound

- Density $p_\zeta(y)$ given by

$$\mathbb{E}_{\mathcal{G}}\left[\frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}(y-G\zeta))\right] = \frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)\mathbb{E}_{\mathcal{G}}\left[\exp(y^\top G\zeta/\sigma^2)\right]$$

Information geometry : the upper bound

- Density $p_\zeta(y)$ given by

$$\mathbb{E}_{\mathcal{G}}\left[\frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}(y - G\zeta))\right] = \frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)\mathbb{E}_{\mathcal{G}}\left[\exp(y^\top G\zeta/\sigma^2)\right]$$

- By Jensen, $p_\theta(y) \geq \frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)$ since $\mathbb{E}_{\mathcal{G}}[G\theta] = 0$

Information geometry : the upper bound

- Density $p_\zeta(y)$ given by

$$\mathbb{E}_{\mathcal{G}}\left[\frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}(y - G\zeta))\right] = \frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)\mathbb{E}_{\mathcal{G}}\left[\exp(y^\top G\zeta/\sigma^2)\right]$$

- By Jensen, $p_\theta(y) \geq \frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)$ since $\mathbb{E}_{\mathcal{G}}[G\theta] = 0$
- $D_{KL}(p_\theta||p_\varphi) \leq \chi^2(\theta, \varphi) = \int \frac{(p_\theta(y) - p_\varphi(y))^2}{p_\theta(y)}\mathrm{d}y$

Information geometry : the upper bound

- Density $p_\zeta(y)$ given by

$$\mathbb{E}_{\mathcal{G}}\left[\frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}(y-G\zeta))\right] = \frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)\mathbb{E}_{\mathcal{G}}\left[\exp(y^\top G\zeta/\sigma^2)\right]$$

- By Jensen, $p_\theta(y) \geq \frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)$ since $\mathbb{E}_{\mathcal{G}}[G\theta] = 0$
- $D_{KL}(p_\theta || p_\varphi) \leq \chi^2(\theta,\varphi) = \int \frac{(p_\theta(y)-p_\varphi(y))^2}{p_\theta(y)}\mathrm{d}y$
- Using $y = G\theta + \sigma\xi$, we can simplify to $\chi^2(\theta,\varphi)$ bounded above by

$$2\mathbb{E}_{\mathcal{G}}\left[\exp(\left(G'\theta\right)^\top G\theta/\sigma^2) - 2\exp(\left(G'\varphi\right)^\top G\theta/\sigma^2) + \exp(\left(G'\varphi\right)^\top G\varphi/\sigma^2)\right]$$

- Density $p_\zeta(y)$ given by

$$\mathbb{E}_{\mathcal{G}}\left[\frac{1}{\sigma^d} g(\sigma^{-1}(y - G\zeta))\right] = \frac{1}{\sigma^d} g(\sigma^{-1} y) \exp(-\|\zeta\|^2/2) \mathbb{E}_{\mathcal{G}}\left[\exp(y^\top G\zeta/\sigma^2)\right]$$

- By Jensen, $p_\theta(y) \geq \frac{1}{\sigma^d} g(\sigma^{-1} y) \exp(-\|\zeta\|^2/2)$ since $\mathbb{E}_{\mathcal{G}}[G\theta] = 0$
- $D_{KL}(p_\theta || p_\varphi) \leq \chi^2(\theta, \varphi) = \int \frac{(p_\theta(y) - p_\varphi(y))^2}{p_\theta(y)} \mathrm{d}y$
- Using $y = G\theta + \sigma\xi$, we can simplify to $\chi^2(\theta, \varphi)$ bounded above by

$$2\mathbb{E}_{\mathcal{G}}\left[\exp((G'\theta)^\top G\theta/\sigma^2) - 2\exp((G'\varphi)^\top G\theta/\sigma^2) + \exp((G'\varphi)^\top G\varphi/\sigma^2)\right]$$

- Expand exponentials to get the upper bound

$$\sum_{m \geq 0} \frac{2}{\sigma^{2m} m!} \mathbb{E}\left[\left((G'\theta)^\top G\theta\right)^m - 2\left((G'\varphi)^\top G\theta\right)^m + \left((G'\varphi)^\top G\varphi\right)^m\right]$$

- Density $p_\zeta(y)$ given by

$$\mathbb{E}_{\mathcal{G}}\left[\frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}(y-G\zeta))\right]=\frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)\mathbb{E}_{\mathcal{G}}\left[\exp(y^\top G\zeta/\sigma^2)\right]$$

- By Jensen, $p_\theta(y)\geq\frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)$ since $\mathbb{E}_{\mathcal{G}}[G\theta]=0$
- $D_{KL}(p_\theta||p_\varphi)\leq\chi^2(\theta,\varphi)=\int\frac{(p_\theta(y)-p_\varphi(y))^2}{p_\theta(y)}\mathrm{d}y$
- Using $y=G\theta+\sigma\xi$, we can simplify to $\chi^2(\theta,\varphi)$ bounded above by

$$2\mathbb{E}_{\mathcal{G}}\left[\exp((G'\theta)^\top G\theta/\sigma^2)-2\exp((G'\varphi)^\top G\theta/\sigma^2)+\exp((G'\varphi)^\top G\varphi/\sigma^2)\right]$$

- Expand exponentials to get the upper bound

$$\sum_{m\geq0}\frac{2}{\sigma^{2m}m!}\mathbb{E}\left[\left((G'\theta)^\top G\theta\right)^m-2\left((G'\varphi)^\top G\theta\right)^m+\left((G'\varphi)^\top G\varphi\right)^m\right]$$

$$=\sum_{m\geq0}\frac{2}{\sigma^{2m}m!}\left\|\mathbb{E}\left[(G\theta)^{\otimes m}\right]\right\|^2-2\left\langle\mathbb{E}\left[(G\theta)^{\otimes m}\right],\mathbb{E}\left[(G\varphi)^{\otimes m}\right]\right\rangle+\left\|\mathbb{E}\left[(G\varphi)^{\otimes m}\right]\right\|^2$$

- Density $p_\zeta(y)$ given by

$$\mathbb{E}_{\mathcal{G}}\left[\frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}(y-G\zeta))\right]=\frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)\mathbb{E}_{\mathcal{G}}\left[\exp(y^\top G\zeta/\sigma^2)\right]$$

- By Jensen, $p_\theta(y)\geq\frac{1}{\sigma^d}\mathrm{g}(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)$ since $\mathbb{E}_{\mathcal{G}}[G\theta]=0$
- $D_{KL}(p_\theta||p_\varphi)\leq\chi^2(\theta,\varphi)=\int\frac{(p_\theta(y)-p_\varphi(y))^2}{p_\theta(y)}\mathrm{d}y$
- Using $y=G\theta+\sigma\xi$, we can simplify to $\chi^2(\theta,\varphi)$ bounded above by

$$2\mathbb{E}_{\mathcal{G}}\left[\exp((G'\theta)^\top G\theta/\sigma^2)-2\exp((G'\varphi)^\top G\theta/\sigma^2)+\exp((G'\varphi)^\top G\varphi/\sigma^2)\right]$$

- Expand exponentials to get the upper bound

$$\sum_{m\geq 0}\frac{2}{\sigma^{2m}m!}\mathbb{E}\left[\left((G'\theta)^\top G\theta\right)^m-2\left((G'\varphi)^\top G\theta\right)^m+\left((G'\varphi)^\top G\varphi\right)^m\right]$$

$$=\sum_{m\geq 0}\frac{2}{\sigma^{2m}m!}\left\|\mathbb{E}\left[(G\theta)^{\otimes m}\right]\right\|^2-2\left\langle\mathbb{E}\left[(G\theta)^{\otimes m}\right],\mathbb{E}\left[(G\varphi)^{\otimes m}\right]\right\rangle+\left\|\mathbb{E}\left[(G\varphi)^{\otimes m}\right]\right\|^2$$

$$=\sum_{m\geq 0}\frac{2}{\sigma^{2m}m!}\|\Delta_m\|^2$$

Information geometry : the upper bound

- Density $p_\zeta(y)$ given by

$$\mathbb{E}_{\mathcal{G}}\left[\frac{1}{\sigma^d}g(\sigma^{-1}(y-G\zeta))\right]=\frac{1}{\sigma^d}g(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)\mathbb{E}_{\mathcal{G}}\left[\exp(y^\top G\zeta/\sigma^2)\right]$$

- By Jensen, $p_\theta(y)\geq\frac{1}{\sigma^d}g(\sigma^{-1}y)\exp(-\|\zeta\|^2/2)$ since $\mathbb{E}_{\mathcal{G}}[G\theta]=0$
- $D_{KL}(p_\theta\|p_\varphi)\leq\chi^2(\theta,\varphi)=\int\frac{(p_\theta(y)-p_\varphi(y))^2}{p_\theta(y)}\mathrm{d}y$
- Using $y=G\theta+\sigma\xi$, we can simplify to $\chi^2(\theta,\varphi)$ bounded above by

$$2\mathbb{E}_{\mathcal{G}}\left[\exp((G'\theta)^\top G\theta/\sigma^2)-2\exp((G'\varphi)^\top G\theta/\sigma^2)+\exp((G'\varphi)^\top G\varphi/\sigma^2)\right]$$

- Expand exponentials to get the upper bound

$$\sum_{m\geq 0}\frac{2}{\sigma^{2m}m!}\mathbb{E}\left[\left((G'\theta)^\top G\theta\right)^m-2\left((G'\varphi)^\top G\theta\right)^m+\left((G'\varphi)^\top G\varphi\right)^m\right]$$

$$=\sum_{m\geq 0}\frac{2}{\sigma^{2m}m!}\left\|\mathbb{E}\left[(G\theta)^{\otimes m}\right]\right\|^2-2\left\langle\mathbb{E}\left[(G\theta)^{\otimes m}\right],\mathbb{E}\left[(G\varphi)^{\otimes m}\right]\right\rangle+\left\|\mathbb{E}\left[(G\varphi)^{\otimes m}\right]\right\|^2$$

$$=\sum_{m\geq 0}\frac{2}{\sigma^{2m}m!}\|\Delta_m\|^2\leq 2\sum_{m=1}^{k-1}\frac{\|\Delta_m\|^2}{\sigma^{2m}m!}+C\cdot\frac{\|\theta\|^{2k-2}\cdot\rho(\theta,\varphi)^2}{\sigma^{2k}}$$

Information geometry : the lower bound

Lemma

Let P_0 *and* P_1 *be any two distributions on a space* $\mathcal{X}$. *If there exists a measurable function* $T : \mathcal{X} \to \mathbb{R}$ *such that* $(\mathbb{E}_0[T(X)] - \mathbb{E}_1[T(X)])^2 = \mu^2$ *and* $\max\{\mathrm{var}_1(T(X)), \mathrm{var}_0(T(X))\} \leq \sigma^2$, *then*

$$D_{KL}(\mathrm{P}_0 \| \mathrm{P}_1) \geq \frac{\mu^2}{4\sigma^2 + \mu^2}$$

Information geometry : the lower bound

Lemma

Let P_0 and P_1 be any two distributions on a space $\mathcal{X}$. If there exists a measurable function $T : \mathcal{X} \to \mathbb{R}$ such that $(\mathbb{E}_0[T(X)] - \mathbb{E}_1[T(X)])^2 = \mu^2$ and $\max\{\mathrm{var}_1(T(X)), \mathrm{var}_0(T(X))\} \leq \sigma^2$, then

$$D_{KL}(\mathrm{P}_0 \| \mathrm{P}_1) \geq \frac{\mu^2}{4\sigma^2 + \mu^2}$$

Corollary

If $\sigma^2 \leq a \cdot \mu$ and $\mu \leq b$ in above, then $D_{KL}(\mathrm{P}_0 \| \mathrm{P}_1) \geq \mu/(4a + b)$.

Lemma

Let P_0 and P_1 be any two distributions on a space $\mathcal{X}$. If there exists a measurable function $T : \mathcal{X} \to \mathbb{R}$ such that $(\mathbb{E}_0[T(X)] - \mathbb{E}_1[T(X)])^2 = \mu^2$ and $\max\{\mathrm{var}_1(T(X)), \mathrm{var}_0(T(X))\} \leq \sigma^2$, then

$$D_{KL}(\mathrm{P}_0 \| \mathrm{P}_1) \geq \frac{\mu^2}{4\sigma^2 + \mu^2}$$

Corollary

If $\sigma^2 \leq a \cdot \mu$ and $\mu \leq b$ in above, then $D_{KL}(\mathrm{P}_0 \| \mathrm{P}_1) \geq \mu/(4a + b)$.

Our goal : To use the Lemma and the Corollary to obtain lower bound on $D_{KL}(p_\theta \| p_\varphi)$.

Lemma

Let P_0 *and* P_1 *be any two distributions on a space* $\mathcal{X}$*. If there exists a measurable function* $T : \mathcal{X} \to \mathbb{R}$ *such that* $(\mathbb{E}_0[T(X)] - \mathbb{E}_1[T(X)])^2 = \mu^2$ *and* $\max\{\mathrm{var}_1(T(X)), \mathrm{var}_0(T(X))\} \leq \sigma^2$*, then*

$$D_{KL}\left(\mathrm{P}_0 \| \mathrm{P}_1\right) \geq \frac{\mu^2}{4\sigma^2 + \mu^2}$$

Corollary

If $\sigma^2 \leq a \cdot \mu$ *and* $\mu \leq b$ *in above, then* $D_{KL}\left(\mathrm{P}_0 \| \mathrm{P}_1\right) \geq \mu/(4a + b)$.

Our goal : To use the Lemma and the Corollary to obtain lower bound on $D_{KL}(p_\theta \| p_\varphi)$.

Need : Suitable statistic T, variance bounds ...

Analysis on Gaussian space

- Let γ be standard Gaussian on $\mathbb{R}$

Analysis on Gaussian space

- Let γ be standard Gaussian on $\mathbb{R}$
- Hermite polynomials in 1 dimension:
  - For $k \geq 0$, the function $h_k(x)$ is a degree- k polynomial.

Analysis on Gaussian space

- Let γ be standard Gaussian on $\mathbb{R}$
- Hermite polynomials in 1 dimension:
  - For $k \geq 0$, the function $h_k(x)$ is a degree- k polynomial.
  - $\{h_k\}_{k\geq 0}$ form an orthogonal basis of of $L_2(\gamma)$

- Let γ be standard Gaussian on $\mathbb{R}$
- Hermite polynomials in 1 dimension:
  - For $k \geq 0$, the function $h_k(x)$ is a degree- k polynomial.
  - $\{h_k\}_{k\geq 0}$ form an orthogonal basis of of $L_2(\gamma)$
  - $\|h_k\|_\gamma^2 = k!$

Analysis on Gaussian space

- Let γ be standard Gaussian on $\mathbb{R}$
- Hermite polynomials in 1 dimension:
  - For $k \geq 0$, the function $h_k(x)$ is a degree- k polynomial.
  - $\{h_k\}_{k\geq 0}$ form an orthogonal basis of of $L_2(\gamma)$
  - $\|h_k\|_\gamma^2 = k!$
  - If $Y \sim \mathcal{N}(\mu, 1)$, then $\mathbb{E}[h_k(Y)] = \mu^k$

Analysis on Gaussian space

- Let γ be standard Gaussian on $\mathbb{R}$
- Hermite polynomials in 1 dimension:
  - For $k \geq 0$, the function $h_k(x)$ is a degree- k polynomial.
  - $\{h_k\}_{k\geq 0}$ form an orthogonal basis of of $L_2(\gamma)$
  - $\|h_k\|_\gamma^2 = k!$
  - If $Y \sim \mathcal{N}(\mu, 1)$, then $\mathbb{E}\left[h_k(Y)\right] = \mu^k$
  - If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}\left[\sigma^k h_k(\sigma^{-1} Y)\right] = \mu^k$
- Hermite polynomials in p dimensions :

Analysis on Gaussian space

- Let γ be standard Gaussian on $\mathbb{R}$
- Hermite polynomials in 1 dimension:
  - For $k \geq 0$, the function $h_k(x)$ is a degree- k polynomial.
  - $\{h_k\}_{k\geq 0}$ form an orthogonal basis of of $L_2(\gamma)$
  - $\|h_k\|_\gamma^2 = k!$
  - If $Y \sim \mathcal{N}(\mu, 1)$, then $\mathbb{E}\left[h_k(Y)\right] = \mu^k$
  - If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}\left[\sigma^k h_k(\sigma^{-1} Y)\right] = \mu^k$
- Hermite polynomials in p dimensions :
  - Given a multi-index $\alpha \in \mathbb{N}^p$, define the multivariate Hermite polynomial h_α by $h_\alpha\left(x_1, \ldots, x_p\right) = \prod_{i=1}^p h_{\alpha_i}\left(x_i\right)$

Analysis on Gaussian space

- Let γ be standard Gaussian on $\mathbb{R}$
- Hermite polynomials in 1 dimension:
    - For $k \geq 0$, the function $h_k(x)$ is a degree- k polynomial.
    - $\{h_k\}_{k\geq 0}$ form an orthogonal basis of of $L_2(\gamma)$
    - $\|h_k\|_\gamma^2 = k!$
    - If $Y \sim \mathcal{N}(\mu, 1)$, then $\mathbb{E}\left[h_k(Y)\right] = \mu^k$
    - If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}\left[\sigma^k h_k(\sigma^{-1} Y)\right] = \mu^k$
- Hermite polynomials in p dimensions :
    - Given a multi-index $\alpha \in \mathbb{N}^p$, define the multivariate Hermite polynomial h_α by $h_\alpha\left(x_1, \ldots, x_p\right) = \prod_{i=1}^p h_{\alpha_i}\left(x_i\right)$
    - The multivariate Hermite polynomials form an orthonormal basis for the space $\mathbb{R}\left[x_1, \ldots, x_p\right]$ of p-variate polynomial functions with respect to the inner product over $L_2\left(\gamma^{\otimes p}\right)$.

Analysis on Gaussian space

- Let γ be standard Gaussian on $\mathbb{R}$
- Hermite polynomials in 1 dimension:
  - For $k \geq 0$, the function $h_k(x)$ is a degree- k polynomial.
  - $\{h_k\}_{k\geq 0}$ form an orthogonal basis of of $L_2(\gamma)$
  - $\|h_k\|_\gamma^2 = k!$
  - If $Y \sim \mathcal{N}(\mu, 1)$, then $\mathbb{E}\left[h_k(Y)\right] = \mu^k$
  - If $Y \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}\left[\sigma^k h_k(\sigma^{-1} Y)\right] = \mu^k$
- Hermite polynomials in p dimensions :
  - Given a multi-index $\alpha \in \mathbb{N}^p$, define the multivariate Hermite polynomial h_α by $h_\alpha\left(x_1, \ldots, x_p\right) = \prod_{i=1}^p h_{\alpha_i}\left(x_i\right)$
  - The multivariate Hermite polynomials form an orthonormal basis for the space $\mathbb{R}\left[x_1, \ldots, x_p\right]$ of p-variate polynomial functions with respect to the inner product over $L_2\left(\gamma^{\otimes p}\right)$.
- In summary, for $Y \sim N_p(\mu, \sigma^2 I_p)$ and $\alpha \in \mathbb{N}^p$, we have $\mathbb{E}\left[\sigma^{\|\alpha\|_1} h_\alpha(\sigma^{-1} Y)\right] = \prod_{i=1}^p \mu_i^{\alpha_i}$.

- Define $H_m(X)$ (for $X \in \mathbb{R}^p$) to be the order m symmetric tensor given by $(H_m(X))_{i_1,\ldots,i_m} = \sigma^m h_\alpha(\sigma^{-1}(X))$. where $\alpha \in \mathbb{N}^p$ is defined by $\alpha_j = |\{k : i_k = j\}|$, for $1 \leq j \leq p$.

The lower bound : constructing the statistic

- Define $H_m(X)$ (for $X \in \mathbb{R}^p$) to be the order m symmetric tensor given by $(H_m(X))_{i_1,\dots,i_m} = \sigma^m h_\alpha(\sigma^{-1}(X))$. where $\alpha \in \mathbb{N}^p$ is defined by $\alpha_j = |\{k : i_k = j\}|$, for $1 \le j \le p$.
- Upshot: if $Y \sim N_p(\mu, \sigma^2 I_p)$, then $(\mathbb{E}[H_m(Y)])_{i_1,\dots,i_m} = \prod_{j=1}^p \mu_j^{\alpha_j}$

- Define $H_m(X)$ (for $X \in \mathbb{R}^p$) to be the order m symmetric tensor given by $(H_m(X))_{i_1,\dots,i_m} = \sigma^m h_\alpha(\sigma^{-1}(X))$. where $\alpha \in \mathbb{N}^p$ is defined by $\alpha_j = |\{k : i_k = j\}|$, for $1 \le j \le p$.
- Upshot: if $Y \sim N_p(\mu, \sigma^2 I_p)$, then $(\mathbb{E}[H_m(Y)])_{i_1,\dots,i_m} = \prod_{j=1}^p \mu_j^{\alpha_j} = \prod_{k=1}^m \mu_{i_k}$

The lower bound : constructing the statistic

- Define $H_m(X)$ (for $X \in \mathbb{R}^p$) to be the order m symmetric tensor given by $(H_m(X))_{i_1,\ldots,i_m} = \sigma^m h_\alpha(\sigma^{-1}(X))$. where $\alpha \in \mathbb{N}^p$ is defined by $\alpha_j = |\{k : i_k = j\}|$, for $1 \leq j \leq p$.
- Upshot: if $Y \sim N_p(\mu, \sigma^2 I_p)$, then $(\mathbb{E}[H_m(Y)])_{i_1,\ldots,i_m} = \prod_{j=1}^p \mu_j^{\alpha_j} = \prod_{k=1}^m \mu_{i_k}$
- In summary, $\mathbb{E}[H_m(Y)] = \mu^{\otimes m}$

The lower bound : constructing the statistic

- Define $H_m(X)$ (for $X \in \mathbb{R}^p$) to be the order m symmetric tensor given by $(H_m(X))_{i_1,\ldots,i_m} = \sigma^m h_\alpha(\sigma^{-1}(X))$. where $\alpha \in \mathbb{N}^p$ is defined by $\alpha_j = |\{k : i_k = j\}|$, for $1 \le j \le p$.
- Upshot: if $Y \sim N_p(\mu, \sigma^2 I_p)$, then $(\mathbb{E}[H_m(Y)])_{i_1,\ldots,i_m} = \prod_{j=1}^p \mu_j^{\alpha_j} = \prod_{k=1}^m \mu_{i_k}$
- In summary, $\mathbb{E}[H_m(Y)] = \mu^{\otimes m}$ (can be used to construct unbiased estimators for $T_k(\theta)$)

The lower bound : constructing the statistic

- For $k \geq 1$, define the degree-k multivariate polynomial on $y = (y_1, \ldots, y_p)$ as:

$$t(y) = \sum_{m=1}^{k} \frac{\langle \Delta_m, H_m(y) \rangle}{(\sqrt{3}\sigma)^{2m} m!}$$

The lower bound : constructing the statistic

- For $k \geq 1$, define the degree-k multivariate polynomial on $y = (y_1, \ldots, y_p)$ as:

$$t(y) = \sum_{m=1}^{k} \frac{\langle \Delta_m, H_m(y) \rangle}{(\sqrt{3}\sigma)^{2m} m!}$$

- If $Y \sim \mathrm{P}_\zeta$, then

$$\mathbb{E}[t(Y)] = \mathbb{E}\left[\sum_{m=1}^{k} \frac{\langle \Delta_m, \mathbb{E}\left[H_m(Y) \mid G\right] \rangle}{(\sqrt{3}\sigma)^{2m} m!}\right] = \sum_{m=1}^{k} \frac{\langle \Delta_m, \mathbb{E}\left[(G\zeta)^{\otimes m}\right] \rangle}{(\sqrt{3}\sigma)^{2m} m!}$$

The lower bound : constructing the statistic

- For $k \geq 1$, define the degree-k multivariate polynomial on $y = (y_1, \ldots, y_p)$ as:

$$t(y) = \sum_{m=1}^{k} \frac{\langle \Delta_m, H_m(y) \rangle}{(\sqrt{3}\sigma)^{2m} m!}$$

- If $Y \sim \mathrm{P}_\zeta$, then

$$\mathbb{E}[t(Y)] = \mathbb{E}\left[\sum_{m=1}^{k} \frac{\langle \Delta_m, \mathbb{E}\left[H_m(Y) \mid G\right] \rangle}{(\sqrt{3}\sigma)^{2m} m!}\right] = \sum_{m=1}^{k} \frac{\left\langle \Delta_m, \mathbb{E}\left[(G\zeta)^{\otimes m}\right] \right\rangle}{(\sqrt{3}\sigma)^{2m} m!}$$

- $\implies \mathbb{E}_{\mathbb{P}_\theta}[t(Y)] - \mathbb{E}_{P_\varphi}[t(Y)]$

$$= \sum_{m=1}^{k} \frac{\left\langle \Delta_m, \left(\mathbb{E}\left[(G\theta)^{\otimes m}\right] - \mathbb{E}\left[(G\varphi)^{\otimes m}\right]\right) \right\rangle}{(\sqrt{3}\sigma)^{2m} m!} = \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m} m!}$$

The lower bound : controlling the variance

- For $Y \sim P_\zeta$, want $\mathrm{Var}[t(Y)] \leq e^{\|\zeta\|^2/\sigma^2} \cdot \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m} m!}$

The lower bound : controlling the variance

- For $Y \sim P_\zeta$, want $\mathrm{Var}[t(Y)] \leq e^{\|\zeta\|^2/\sigma^2} \cdot \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m} m!}$
- If $Z \sim P_0$, then

$$\mathbb{E}[t(Y)^2] \leq \mathbb{E}\left[t(Z)^2 \cdot \frac{\mathrm{d}P_\zeta}{\mathrm{d}P_0}\right] \leq (\mathbb{E}[t(Z)^4])^{1/2}(\chi^2(P_\zeta, P_0)+1)^{1/2}$$

The lower bound : controlling the variance

- For $Y \sim P_\zeta$, want $\mathrm{Var}[t(Y)] \leq e^{\|\zeta\|^2/\sigma^2} \cdot \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3}\sigma)^{2m} m!}$
- If $Z \sim P_0$, then

$$\mathbb{E}[t(Y)^2] \leq \mathbb{E}\left[t(Z)^2 \cdot \frac{\mathrm{d}P_\zeta}{\mathrm{d}P_0}\right] \leq (\mathbb{E}[t(Z)^4])^{1/2}(\chi^2(P_\zeta, P_0)+1)^{1/2}$$

- $\chi^2(P_\zeta, P_0) + 1 \leq e^{\|\zeta\|^2/\sigma^2}$ (direct computation)

Controlling the variance : Gaussian hypercontractivity

- For $Z \sim P_0$, want $\mathbb{E}[t(Z)^4]^{1/2} \leq \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m} \sigma^{2m} m!}$

Controlling the variance : Gaussian hypercontractivity

- For $Z \sim P_0$, want $\mathbb{E}[t(Z)^4]^{1/2} \leq \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m}\sigma^{2m} m!}$
- Gaussian noise operator & Ornstein-Uhlenbeck process :
  $[U_\rho f](x) = \mathbb{E}_{g \sim N(0,I)}\left[f\left(\rho x + \sqrt{1-\rho^2}\, g\right)\right]$

Controlling the variance : Gaussian hypercontractivity

- For $Z \sim P_0$, want $\mathbb{E}[t(Z)^4]^{1/2} \leq \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m}\sigma^{2m} m!}$
- Gaussian noise operator & Ornstein-Uhlenbeck process :
  $[U_\rho f](x) = \mathbb{E}_{g\sim N(0,I)}\left[f\left(\rho x + \sqrt{1-\rho^2}\, g\right)\right]$
- Gaussian Hypercontractivity : For
  $1 \leq p \leq q \leq \infty, \|U_\rho f\|_q \leq \|f\|_p \forall\ 0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$ in Gaussian space

Controlling the variance : Gaussian hypercontractivity

- For $Z \sim P_0$, want $\mathbb{E}[t(Z)^4]^{1/2} \leq \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m} \sigma^{2m} m!}$
- Gaussian noise operator & Ornstein-Uhlenbeck process : $[U_\rho f](x) = \mathbb{E}_{g \sim N(0,I)} \left[f\left(\rho x + \sqrt{1-\rho^2}\, g\right) \right]$
- Gaussian Hypercontractivity : For $1 \leq p \leq q \leq \infty, \|U_\rho f\|_q \leq \|f\|_p \forall\ 0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$ in Gaussian space
- $U_\rho h_k = \rho^k h_k$ (in 1D);

Controlling the variance : Gaussian hypercontractivity

- For $Z \sim P_0$, want $\mathbb{E}[t(Z)^4]^{1/2} \leq \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m} \sigma^{2m} m!}$
- Gaussian noise operator & Ornstein-Uhlenbeck process : $[U_\rho f](x) = \mathbb{E}_{g \sim N(0,I)}\left[f\left(\rho x + \sqrt{1-\rho^2}\, g\right)\right]$
- Gaussian Hypercontractivity : For $1 \leq p \leq q \leq \infty, \|U_\rho f\|_q \leq \|f\|_p \forall\ 0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$ in Gaussian space
- $U_\rho h_k = \rho^k h_k$ (in 1D); $U_\rho h_\alpha = \rho^{\|\alpha\|_1} h_\alpha$ (in general)

Controlling the variance : Gaussian hypercontractivity

- For $Z \sim P_0$, want $\mathbb{E}[t(Z)^4]^{1/2} \leq \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m}\sigma^{2m} m!}$
- Gaussian noise operator & Ornstein-Uhlenbeck process : $[U_\rho f](x) = \mathbb{E}_{g\sim N(0,I)}\left[f\left(\rho x + \sqrt{1-\rho^2}\, g\right)\right]$
- Gaussian Hypercontractivity : For $1 \leq p \leq q \leq \infty, \|U_\rho f\|_q \leq \|f\|_p \forall\ 0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$ in Gaussian space
- $U_\rho h_k = \rho^k h_k$ (in 1D); $U_\rho h_\alpha = \rho^{\|\alpha\|_1} h_\alpha$ (in general)
- Define polynomial $\tilde{t}(y) = \sum_{m=1}^{k} \frac{\langle \Delta_m, H_m(y)\rangle}{(\sqrt{3})^{m}\sigma^{2m} m!}$

Controlling the variance : Gaussian hypercontractivity

- For $Z \sim P_0$, want $\mathbb{E}[t(Z)^4]^{1/2} \leq \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m} \sigma^{2m} m!}$
- Gaussian noise operator & Ornstein-Uhlenbeck process : $[U_\rho f](x) = \mathbb{E}_{g \sim N(0,I)} \left[f\left(\rho x + \sqrt{1-\rho^2}\, g\right) \right]$
- Gaussian Hypercontractivity : For $1 \leq p \leq q \leq \infty, \|U_\rho f\|_q \leq \|f\|_p \forall\, 0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$ in Gaussian space
- $U_\rho h_k = \rho^k h_k$ (in 1D); $U_\rho h_\alpha = \rho^{\|\alpha\|_1} h_\alpha$ (in general)
- Define polynomial $\tilde{t}(y) = \sum_{m=1}^{k} \frac{\langle \Delta_m, H_m(y) \rangle}{(\sqrt{3})^m \sigma^{2m} m!}$
- Observe that $t = U_{1/\sqrt{3}} \tilde{t}$ as functions

Controlling the variance : Gaussian hypercontractivity

- For $Z \sim P_0$, want $\mathbb{E}[t(Z)^4]^{1/2} \le \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m}\sigma^{2m} m!}$
- Gaussian noise operator & Ornstein-Uhlenbeck process : $[U_\rho f](x) = \mathbb{E}_{g\sim N(0,I)}\left[f\left(\rho x + \sqrt{1-\rho^2}\, g\right)\right]$
- Gaussian Hypercontractivity : For $1 \le p \le q \le \infty, \|U_\rho f\|_q \le \|f\|_p \forall\ 0 \le \rho \le \sqrt{\frac{p-1}{q-1}}$ in Gaussian space
- $U_\rho h_k = \rho^k h_k$ (in 1D); $U_\rho h_\alpha = \rho^{\|\alpha\|_1} h_\alpha$ (in general)
- Define polynomial $\tilde{t}(y) = \sum_{m=1}^{k} \frac{\langle \Delta_m, H_m(y)\rangle}{(\sqrt{3})^m \sigma^{2m} m!}$
- Observe that $t = U_{1/\sqrt{3}}\tilde{t}$ as functions
- Gaussian Hypercontractivity : in Gaussian space, we have $\|t\|_4 \le \|\tilde{t}\|_2$

Controlling the variance : Gaussian hypercontractivity

- For $Z \sim P_0$, want $\mathbb{E}[t(Z)^4]^{1/2} \leq \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m}\sigma^{2m} m!}$
- Gaussian noise operator & Ornstein-Uhlenbeck process :
  $[U_\rho f](x) = \mathbb{E}_{g\sim N(0,I)}\left[f\left(\rho x + \sqrt{1-\rho^2}\, g\right)\right]$
- Gaussian Hypercontractivity : For
  $1 \leq p \leq q \leq \infty, \|U_\rho f\|_q \leq \|f\|_p \forall\ 0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$ in Gaussian space
- $U_\rho h_k = \rho^k h_k$ (in 1D); $U_\rho h_\alpha = \rho^{\|\alpha\|_1} h_\alpha$ (in general)
- Define polynomial $\tilde{t}(y) = \sum_{m=1}^{k} \frac{\langle \Delta_m, H_m(y)\rangle}{(\sqrt{3})^m \sigma^{2m} m!}$
- Observe that $t = U_{1/\sqrt{3}}\tilde{t}$ as functions
- Gaussian Hypercontractivity : in Gaussian space, we have
  $\|t\|_4 \leq \|\tilde{t}\|_2 \iff \mathbb{E}[t(Z)^4]^{1/4} \leq \mathbb{E}[t(Z)^2]^{1/2}$

- For $Z \sim P_0$, want $\mathbb{E}[t(Z)^4]^{1/2} \leq \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m}\sigma^{2m}m!}$
- Gaussian noise operator & Ornstein-Uhlenbeck process : $[U_\rho f](x) = \mathbb{E}_{g\sim N(0,I)}\left[f\left(\rho x + \sqrt{1-\rho^2}\, g\right)\right]$
- Gaussian Hypercontractivity : For $1 \leq p \leq q \leq \infty, \|U_\rho f\|_q \leq \|f\|_p \forall\ 0 \leq \rho \leq \sqrt{\frac{p-1}{q-1}}$ in Gaussian space
- $U_\rho h_k = \rho^k h_k$ (in 1D); $U_\rho h_\alpha = \rho^{\|\alpha\|_1} h_\alpha$ (in general)
- Define polynomial $\tilde{t}(y) = \sum_{m=1}^{k} \frac{\langle \Delta_m, H_m(y)\rangle}{(\sqrt{3})^m \sigma^{2m} m!}$
- Observe that $t = U_{1/\sqrt{3}}\tilde{t}$ as functions
- Gaussian Hypercontractivity : in Gaussian space, we have $\|t\|_4 \leq \|\tilde{t}\|_2 \iff \mathbb{E}[t(Z)^4]^{1/4} \leq \mathbb{E}[t(Z)^2]^{1/2}$
- Explicit computation : $\mathbb{E}[t(Z)^2] = \sum_{m=1}^{k} \frac{\|\Delta_m\|^2}{(\sqrt{3})^{2m}\sigma^{2m}m!}$

References

• "Sparse Multi-Reference Alignment: Phase Retrieval, Uniform Uncertainty Principles and the Beltway Problem." G. and Rigollet, Foundations of Computational Mathematics (2023).

• "Dictionary Learning under Symmetries via Group Representations.", G., Low, Soh, Feng and Tan, arXiv preprint arXiv:2305.19557.

• "Minimax-optimal estimation for sparse multi-reference alignment with collision-free signals.", G., Mukherjee and Pan, arXiv preprint arXiv:2312.07839.

• "Likelihood landscape and maximum likelihood estimation for the discrete orbit recovery model." Fan, Sun, Wang and Wu, Communications on Pure and Applied Mathematics (2020).

• "Estimation under group actions: recovering orbits from invariants." Bandeira, Blum-Smith, Kileel, Perry, Weed and Wein, Applied and Computational Harmonic Analysis (2023)

• "The sample complexity of multireference alignment." Perry, Weed, Bandeira, Rigollet and Singer, SIAM Journal on Mathematics of Data Science (2019).

• "Optimal rates of estimation for multi-reference alignment." Bandeira, Niles-Weed and Rigollet, Mathematical Statistics and Learning (2020).