A Random CSP with Connections to Discrepancy Theory and Randomized Trials

Eren C. Kızıldağ (University of Illinois Urbana-Champaign)

International Symposium on Information Theory

July 12, 2024

Overview of Model

Fix $d \in \mathbb{N}$ and $\boldsymbol{c} = (c_1, \dots, c_d) \in \mathbb{R}_+^d$. Generate **iid** $X_1, \dots, X_d \sim \mathcal{N}(0, I_n)$. Define

$$\mathcal{F}(\boldsymbol{c}) = \left\{ \boldsymbol{\sigma} \in \{-1, 1\}^n : |\langle \boldsymbol{\sigma}, X_i \rangle| \le \sqrt{n} 2^{-c_i n}, \forall i \right\}.$$

Focus: Non-proportional regime, $n \to \infty, d = O(1)$. **Random CSP**

Questions: When is $\mathcal{F}(\boldsymbol{c}) \ne \varnothing$? How does its 'geometry' look like?

Today: Sharp Phase Transition for $\{\mathcal{F}(\boldsymbol{c}) \ne \varnothing\}$. Landscape of $\mathcal{F}(\boldsymbol{c})$

Motivation

Discrepancy Theory

- Given $M \in \mathbb{R}^{d \times n}$, compute/bound its **discrepancy** $\mathcal{D}(M) := \min_{\boldsymbol{\sigma} \in \{\pm 1\}^n} \|M\boldsymbol{\sigma}\|_\infty$.
- Note that $\mathcal{D}(M) \le B \iff \exists \boldsymbol{\sigma} \in \{\pm 1\}^n : |\langle \boldsymbol{\sigma}, X_i \rangle| \le B, \forall i$ (X_i are rows of M).
- **Worst-case** & **random** M. **Existential** & **Algorithmic** results.
  [Spencer 85, Karmarkar-Karp-Lueker-Odlyzko 86, Matousek 99, Chazelle 00, Bansal 10, Lovett-Meka 15]

Proportional Regime: For $d = \Theta(n)$, and $M \in \mathbb{R}^{d \times n}$ with $\mathcal{N}(0,1)$ entries, whp

$$\mathcal{D}(M) = f(\alpha)\sqrt{n}(1 + o_n(1)), \quad \text{where} \quad \alpha = d/n$$

for explicit $f(\cdot)$. [Perkins-Xu 21, Abbe-Li-Sly 21]

Motivation: Discrepancy and Random CSPs

Symmetric Binary Perceptron

Fix $\kappa > 0$ and consider random $M \in \mathbb{R}^{\alpha n \times n}$. What is the largest $\alpha > 0$ for which a

$$\boldsymbol{\sigma} \in \{\pm 1\}^n : \|M\boldsymbol{\sigma}\|_\infty \le \kappa\sqrt{n}$$

exists whp? When do efficient search algs work?

- **Perceptron:** Model for pattern storage. Popular in **probability**, **stat phys**, **statistics** communities [Cover 65, Hopfield 82, Krauth-Mézard 89, Talagrand 99, 10, Franz-Parisi 16, Candès-Sur 20, Perkins-Xu 21, Abbe-Li-Sly 21, 22, Montanari-Zhong-Zhou 21, Sah-Sawhney 23, Nakajima-Sun 23, Barbier-El Alaoui-Krzakala-Zdeborovà 23, Gamarnik-**K.**-Perkins-Xu 22, 23, **K.**-Wakhare 23]
- **Dual** of **discrepancy**: Fix $\alpha > 0$, seek smallest $\kappa > 0$ s.t. $\boldsymbol{\sigma} \in \{\pm 1\}^n : \|M\boldsymbol{\sigma}\|_\infty \le \kappa\sqrt{n}$ exists

Motivation: Our Model

$M \in \mathbb{R}^{d\times n}$ with **iid** rows $X_1,\dots,X_d \sim \mathcal{N}(0,I_n)$, $\boldsymbol{c} = (c_1,\dots,c_d) \in \mathbb{R}_+^d$.

$$\mathcal{F}(\boldsymbol{c}) = \left\{\boldsymbol{\sigma} \in \{-1,1\}^n : |\langle \boldsymbol{\sigma}, X_i\rangle| \le \sqrt{n}2^{-c_i n}, \forall i\right\}.$$

$$\boldsymbol{\sigma} \in \mathcal{F}(\boldsymbol{c}) \iff \left|M\boldsymbol{\sigma}\right| \le \begin{pmatrix} \sqrt{n}2^{-c_1 n} \\ \vdots \\ \sqrt{n}2^{-c_d n}\end{pmatrix}.$$

- Dual of discrepancy. **Non-proportional** regime, $d = O_n(1)$. **Non-uniform** constraints
- 2^{-n} scaling: $\min_{\boldsymbol{\sigma}\in\{\pm 1\}^n} \|M\boldsymbol{\sigma}\|_\infty = \sqrt{n}2^{-\Omega(n/d)} = \sqrt{n}2^{-\Omega(n)}$ for $d = O_n(1)$
  [Karmarkar-Karp-Lueker-Odlyzko 86, Costello 09, Turner-Meka-Rigollet 20]

Motivation: Randomized Controlled Trials

- Gold standard for clinical trials (drug/vaccine)
- n individuals, covariates $Y_1, \dots, Y_n \in \mathbb{R}^d$ (columns of $M \in \mathbb{R}^{d \times n}$). $n \gg d$.
- Split into balanced **treatment** & **control**: for thresholds $t_1, \dots, t_d$ and features $j \in \{1, \dots, d\}$

$$\mathcal{D}_j := \left| \sum_{i:\boldsymbol{\sigma}(i)=+1} Y_i(j) - \sum_{i:\boldsymbol{\sigma}(i)=-1} Y_i(j) \right| \le t_j, \quad \forall j \in [d].$$

- Any solution to **CSP** gives a valid design: $\boldsymbol{\sigma} \in \mathcal{F}(\boldsymbol{c}) \Leftrightarrow D_j \le t_j, \quad \text{for} \quad t_j = \sqrt{n} 2^{-c_j n}$.

Main Results: A Sharp Phase Transition

For $d \in \mathbb{N}, \boldsymbol{c} = (c_1, \dots, c_d) \in \mathbb{R}_+^d$ and **iid** $X_1, \dots, X_d \sim \mathcal{N}(0, I_n)$

$$\mathcal{F}(\boldsymbol{c}) = \{\boldsymbol{\sigma} \in \{-1,1\}^n : |\langle \boldsymbol{\sigma}, X_i \rangle| \le \sqrt{n} 2^{-c_i n}, \forall i\}$$

Theorem (**K.**, 2024)

$$\lim_{n \to \infty} \mathbb{P}\left[\mathcal{F}(\boldsymbol{c}) \ne \varnothing\right] = \begin{cases} 0, & \text{if } \|\boldsymbol{c}\|_1 > 1 \\ 1, & \text{if } \|\boldsymbol{c}\|_1 < 1. \end{cases}$$

- [Costello 09]: $\min_{\boldsymbol{\sigma} \in \{\pm 1\}^n} \|M\boldsymbol{\sigma}\|_\infty \sim \sqrt{n} 2^{-n/d}$ for $M \in \mathbb{R}^{d \times n}, d = O(1)$. ($c_i \sim 1/d$)
- Proof based on the **first moment method** and the **second moment method**

Proof Sketch: $\|c\|_1 > 1$

Let $T = |\mathcal{F}(\boldsymbol{c})|$. Our proof is based on the **moment method**.

First Moment Method for $\|c\|_1 > 1$

Observe that $n^{-1/2}\langle \boldsymbol{\sigma}, X_i \rangle \sim \mathcal{N}(0,1), 1 \le i \le d$ are **iid**. So,

$$\mathbb{P}[\boldsymbol{\sigma} \in \mathcal{F}(\boldsymbol{c})] = \prod_{1 \le i \le d} \mathbb{P}[|\mathcal{N}(0,1)| \le 2^{-c_i n}] \sim 2^{-\|\boldsymbol{c}\|_1 n}.$$

Using **Markov's inequality**, we have that for $\|\boldsymbol{c}\|_1 > 1$

$$\mathbb{P}[T \ge 1] \le \mathbb{E}[T] \sim 2^{n(1-\|\boldsymbol{c}\|_1)} = 2^{-\Theta(n)}.$$

Hence, $\mathcal{F}(\boldsymbol{c}) = \varnothing$ whp for $\|\boldsymbol{c}\|_1 > 1$.

Proof Idea: $\|c\|_1 < 1$

Paley-Zygmund Inequality (Second Moment Method)

Let $T \ge 0$ be a rv and $\theta \in [0,1]$. Then,

$$\mathbb{P}\big[T > \theta \mathbb{E}[T]\big] \ge (1-\theta)^2 \frac{\mathbb{E}[T]^2}{\mathbb{E}[T^2]}.$$

Suppose $T \in \mathbb{Z}$ and $\mathbb{E}[T^2] = (1+o(1))\mathbb{E}[T]^2$. Taking $\theta = 0$ yields $T \ge 1$ whp.

$$\mathbb{E}[T^2] = \underbrace{\sum_{(\boldsymbol{\sigma},\boldsymbol{\sigma}')\in\mathcal{T}_1} \mathbb{P}[\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \mathcal{F}(\boldsymbol{c})]}_{:=\Sigma_1} + \underbrace{\sum_{(\boldsymbol{\sigma},\boldsymbol{\sigma}')\in\mathcal{T}_2} \mathbb{P}[\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \mathcal{F}(\boldsymbol{c})]}_{:=\Sigma_2},$$

where for an arbitrary $\epsilon > 0$,

$$\mathcal{T}_1 = \left\{(\boldsymbol{\sigma},\boldsymbol{\sigma}') : \frac{1}{n}\langle \boldsymbol{\sigma},\boldsymbol{\sigma}'\rangle \in [-\epsilon,\epsilon]\right\} \quad \text{and} \quad \mathcal{T}_2 = \left\{(\boldsymbol{\sigma},\boldsymbol{\sigma}') : \frac{1}{n}|\langle \boldsymbol{\sigma},\boldsymbol{\sigma}'\rangle| > \epsilon\right\}$$

Proof Sketch: $\|c\|_1 < 1$

- Show $\Sigma_2 \le \mathbb{E}[T]^2 e^{-\Theta(n)}$. For Σ_1, take $(\boldsymbol{\sigma}, \boldsymbol{\sigma}') \in \mathcal{T}_1$. Then,

$$\left(\frac{1}{\sqrt{n}}\langle \boldsymbol{\sigma}, X_i\rangle, \frac{1}{\sqrt{n}}\langle \boldsymbol{\sigma}', X_i\rangle\right) \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad \text{where} \quad \rho = \frac{1}{n}\langle \boldsymbol{\sigma}, \boldsymbol{\sigma}'\rangle \in [-\epsilon, \epsilon]$$

- Ignoring absolute constants (not depending on ϵ)

$$\mathbb{P}\left[\boldsymbol{\sigma}, \boldsymbol{\sigma}' \in \mathcal{F}(\boldsymbol{c})\right] \le \left(1-\epsilon^2\right)^{-\frac{d}{2}} 2^{-2\|\boldsymbol{c}\|_1 n}$$

- This gives $\Sigma_1 \le \mathbb{E}[T]^2 \left(1-\epsilon^2\right)^{-\frac{d}{2}}$. Using $d = O(1)$,

$$\liminf_{n\to\infty} \mathbb{P}[T \ge 1] \ge \liminf_{n\to\infty} \frac{\mathbb{E}[T]^2}{\mathbb{E}[T^2]} \ge (1-\epsilon^2)^{\frac{d}{2}}.$$

- Conclude by $\epsilon \to 0$

Main Results: Solution Space Geometry

$$\mathcal{F}(\boldsymbol{c}) = \left\{\boldsymbol{\sigma} \in \{-1,1\}^n : |\langle \boldsymbol{\sigma}, X_i\rangle| \le \sqrt{n}2^{-c_i n}, \forall i \in [d]\right\}.$$

Theorem (K., 2024)

Let $\|\boldsymbol{c}\|_1 > \frac{1}{2}$. There exists $\beta^* \in (0,1)$ such that whp

$$\min_{\boldsymbol{\sigma},\boldsymbol{\sigma}' \in \mathcal{F}(\boldsymbol{c}), \boldsymbol{\sigma} \ne \boldsymbol{\sigma}'} d_H(\boldsymbol{\sigma}, \boldsymbol{\sigma}') \ge \beta^* n.$$

- **Solutions are Isolated**: If $\|\boldsymbol{c}\|_1 > \frac{1}{2}$ then any $(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ are $\Omega(n)$ apart.
- Proof via **first moment method**: let T count $\#(\boldsymbol{\sigma}, \boldsymbol{\sigma}') : d_H(\boldsymbol{\sigma}, \boldsymbol{\sigma}') \le \beta^* n$, show $\mathbb{E}[T] = o(1)$.
- Suggests **algorithmic hardness**
  [Achlioptas-Ricci Tersenghi 06, Achlioptas-Coja Oghlan 08, Gamarnik-Sudan 14, 17, Gamarnik-**K.** 21]

Main Results: Solution Space Geometry

Theorem (K., 2024)

Let $\|\boldsymbol{c}\|_1 < \frac{1}{2}$ *and* $\beta \in (0,1)$ *be arbitrary. Then,*

$$\mathbb{E}[N_\beta] = e^{\Theta(n)}, \quad \textit{where} \quad N_\beta = \left|(\boldsymbol{\sigma}, \boldsymbol{\sigma}') : \boldsymbol{\sigma}, \boldsymbol{\sigma}' \in \mathcal{F}(\boldsymbol{c}), 1 \le d_H(\boldsymbol{\sigma}, \boldsymbol{\sigma}') \le \beta n\right|.$$

- **First moment evidence** that $\exists(\boldsymbol{\sigma}, \boldsymbol{\sigma}')$ at arbitrarily small distances for $\|\boldsymbol{c}\|_1 < \frac{1}{2}$.
- Matching **second moment bound:** Can we show

$$\mathbb{E}[N_\beta^2] = (1 + o(1))\mathbb{E}[N_\beta]^2$$

and get $N_\beta \ge 1$ via **Paley-Zygmund**?

Independent Instances

$\mathcal{F}(\boldsymbol{c})$ defined before. For $\boldsymbol{c}' = (c_1', \dots, c_d') \in \mathbb{R}_+^d$ and **iid** $X_1', \dots, X_d' \sim \mathcal{N}(0, I_n)$, let

$$\mathcal{F}'(\boldsymbol{c}') = \{\boldsymbol{\sigma} \in \{-1,1\}^n : |\langle \boldsymbol{\sigma}, X_i' \rangle| \le \sqrt{n} 2^{-c_i' n}, \forall i \in [d]\}.$$

When is $\mathcal{F}(\boldsymbol{c}) \cap \mathcal{F}'(\boldsymbol{c}') \ne \varnothing$? If $\cap$ is empty, how far $\mathcal{F}(\boldsymbol{c})$ and $\mathcal{F}'(\boldsymbol{c}')$ are?

Motivation

- RCT $\boldsymbol{\sigma} \in \mathcal{F}(\boldsymbol{c})$. Design a new RCT $\boldsymbol{\sigma}'$ involving a new population & different constraints $\boldsymbol{c}'$.
- Repeat similar **RCT** at different regions or many years later: populations **do not** overlap

Can the same RCT $\boldsymbol{\sigma}$ be used as is? If not, how many changes are needed?

Solutions Spaces of Independent Instances

When is $\mathcal{F}(\boldsymbol{c}) \cap \mathcal{F}(\boldsymbol{c}') \neq \varnothing$?

Consider $\bar{\boldsymbol{c}} = (\boldsymbol{c}, \boldsymbol{c}') \in \mathbb{R}_+^{2d}$ and **iid** $X_1, \dots, X_d, X_1', \dots, X_d' \sim \mathcal{N}(0, I_n)$. We immediately obtain

Corollary (to Theorem 1)

$\mathcal{F}(\boldsymbol{c}) \cap \mathcal{F}'(\boldsymbol{c}') \neq \varnothing$ whp if $\|\boldsymbol{c}\|_1 + \|\boldsymbol{c}'\|_1 < 1$ and $\mathcal{F}(\boldsymbol{c}) \cap \mathcal{F}'(\boldsymbol{c}') = \varnothing$ whp if $\|\boldsymbol{c}\|_1 + \|\boldsymbol{c}'\|_1 > 1$.

Suppose $\|\boldsymbol{c}\|_1 + \|\boldsymbol{c}'\|_1 > 1$. How far $\mathcal{F}(\boldsymbol{c})$ and $\mathcal{F}'(\boldsymbol{c}')$ are?

Main Results: Distance between Independent Instances

Let $\|\boldsymbol{c}\|_1 + \|\boldsymbol{c}'\|_1 > 1$ and $d(\boldsymbol{c}, \boldsymbol{c}') := \min_{\boldsymbol{\sigma} \in \mathcal{F}(\boldsymbol{c}), \boldsymbol{\sigma}' \in \mathcal{F}'(\boldsymbol{c}')} \frac{d_H(\boldsymbol{\sigma}, \boldsymbol{\sigma}')}{n}$, where d_H is Hamming distance.

Theorem (K., 2024)

Suppose $\gamma^* \in (0, \frac{1}{2})$ is the unique value such that

$$h(\gamma^*) = \|\boldsymbol{c}\|_1 + \|\boldsymbol{c}'\|_1 - 1, \quad \text{where} \quad h(p) = -p\log_2 p - (1-p)\log_2(1-p).$$

Then, for any $\epsilon > 0$, $\lim_{n\to\infty}\left[|d(\boldsymbol{c}, \boldsymbol{c}') - \gamma^*| \le \epsilon\right] = 1$.

- $d(\boldsymbol{c}, \boldsymbol{c}') \xrightarrow{\text{i.p.}} \gamma^*$, which is well-defined as $h : [0, \frac{1}{2}] \to [0, 1]$ is a bijection.
- $\mathcal{F}(\boldsymbol{c})$ and $\mathcal{F}'(\boldsymbol{c}')$ are $\Omega(n)$ apart.

Future Work

- **Universality**: Is **Gaussianity** necessary?
- Second moment calculation for N_β
- **Algorithmic** guarantees: Can we find a $\boldsymbol{\sigma} \in \mathcal{F}(\boldsymbol{c})$ in poly time?
- What are the **fundamental limits** of algs? **Overlap Gap Property**
  [Gamarnik-Sudan 14, 17, Gamarnik-Jagannath-Wein 20, Gamarnik-**K.** 21, Gamarnik-**K.**-Perkins-Xu 22,23]

Thank you!