

# THINNING TO IMPROVE TWO-SAMPLE DISCREPANCY

GLEB SMIRNOV AND ROMAN VERSHYNIN

ABSTRACT. The discrepancy between two independent samples  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_n$  drawn from the same distribution on  $\mathbb{R}^d$  typically has order  $O(\sqrt{n})$  even in one dimension. We give a simple online algorithm that reduces the discrepancy to  $O(\log^{2d} n)$  by discarding a small fraction of the points.

Random sampling inevitably brings errors. Classical work in statistics and probability has led to a thorough understanding of sampling errors. Much less is known about how to reduce them with minimal intervention. In this paper, we address the following version of the problem – how to align two independent random samples with each other by discarding a few points?

**1. Main results.** Let  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_n$  be two independent samples, each consisting of  $n$  independent observations drawn from the same Borel probability distribution on  $\mathbb{R}^d$ . For a Borel set  $B \subset \mathbb{R}^d$ , we count how many points from each sample fall into  $B$ , and define its discrepancy as:

$$\text{disc}(B) := |\#\{i : X_i \in B\} - \#\{i : Y_i \in B\}|.$$

Taking the supremum over all axis-aligned, half-infinite<sup>1</sup> boxes  $B = (-\infty, b_1] \times \dots \times (-\infty, b_d]$ , define:

$$D_{n,n} := \sup_B \text{disc}(B). \tag{1.1}$$

In dimension  $d = 1$ , the normalized discrepancy  $n^{-1}D_{n,n}$  recovers the Kolmogorov-Smirnov distance. Classical VC theory [15, Theorem 8.3.23] gives:

$$\mathbb{E} D_{n,n} = O_d(\sqrt{n}).$$

This order is optimal for any nontrivial distribution, because we have  $\mathbb{E} \text{disc}(B) \asymp \sqrt{n}$  for any set  $B$  whose measure is bounded away from 0 and 1. However, the discrepancy can be drastically reduced by discarding a small fraction of the samples:

**Theorem 1** (Two-sample discrepancy). *Fix  $T$  so that  $1 \leq T \leq \sqrt{n}$ . Let  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_n$  be i.i.d. samples from the same Borel probability distribution on  $\mathbb{R}^d$ . There is a randomized online algorithm that discards, on average, at most  $Cn/T$  of the  $X_i$ 's and  $Y_i$ 's, and achieves*

$$\mathbb{E} D_{n,n} \leq T \log_2^{2d} n,$$

where  $C$  is an absolute constant. Expectations are over both samples and the algorithm.

The algorithm (described in §3) does not need to know the distribution. It is online in the sense that it processes points one by one, with the location of each point revealed only when processed. The decision to keep or discard a point depends only on its value and past decisions, and is final. At each step, the algorithm randomly chooses whether to process the next point from the  $X$ - or

---

R.V. is partially supported by NSF Grant DMS 1954233 and NSF+Simons Research Collaborations on the Mathematical and Scientific Foundations of Deep Learning.

<sup>1</sup>Our main results remain valid if the supremum is taken over all finite boxes  $B = [a_1, b_1] \times \dots \times [a_d, b_d]$ .

$Y$ -sample by tossing a fair  $\pm 1$  coin: if  $+1$ , a point from  $X$ ; if  $-1$ , from  $Y$ . The procedure stops when one sample is exhausted; all remaining points from the other are discarded. On average,  $O(\sqrt{n})$  points are left unprocessed, which is absorbed into the  $O(n/T)$  discard budget.

Let  $x_1, \dots, x_k \in \mathbb{R}^d$ , with  $k < 2n$ , denote the sequence of processed points. For each  $i$ , define  $\varepsilon_i = 1$  if  $x_i$  is from the  $X$ -sample and  $\varepsilon_i = -1$  if  $x_i$  is from the  $Y$ -sample. Then  $\varepsilon_1, \dots, \varepsilon_k$  are i.i.d. Rademacher random variables. For any Borel set  $B \subset \mathbb{R}^d$ , define the *sign discrepancy*:

$$\text{disc}_k(B) = \left| \sum_{i=1}^k \varepsilon_i \mathbf{1}_{\{x_i \in B\}} \right|. \quad (1.2)$$

This reduces Theorem 1 to:

**Proposition 1** (Sign discrepancy). *Let  $m$  be an integer with  $m \leq 2n$ . Let  $x_1, \dots, x_m$  be i.i.d. samples from a Borel probability distribution in  $\mathbb{R}^d$ , let  $\varepsilon_1, \dots, \varepsilon_m$  be independent Rademacher random variables, and let  $T > 0$ . There is an online algorithm that discards, on average, at most  $n/T$  of the  $x_i$ 's, and achieves*

$$\mathbb{E} \sup_{B, 1 \leq k \leq m} \text{disc}_k(B) \leq T \log_2^{2d} n + d \ln(dn) + C_1 d,$$

where the supremum is over all axis-aligned, half-infinite boxes  $B$  and all  $k = 1, \dots, m$ , and where  $C_1$  is an absolute constant.

Specializing to  $m = 2n$  recovers Theorem 1.

*Remark 1* (Time and memory cost). The proof shows that each point can be processed in time  $O(\log_2^d n)$ , independent of the parameter  $T$ , yielding a total time complexity of  $O(n \log_2^d n)$ . The algorithm also requires  $O(n \log_2^d n)$  memory.

*Remark 2* (Prior work). Constructing point sets with low discrepancy (star and two-color) is a classical, well-studied problem; see the monographs [13, 10, 7]. Less is known about how to *improve* the discrepancy of a given random sample by moving or discarding only a few points. Recent progress appears in [12, 9, 11, 8, 14].

Closest to our setting is the work [12], which considers the star discrepancy – a stronger notion than the two-sample discrepancy – that compares the empirical measure with the true measure over axis-aligned boxes. The authors provide an online algorithm for points sampled from the uniform distribution on  $[0, 1]^d$ ; their algorithm discards  $O(n/T)$  points and achieves star discrepancy  $O(T \log_2^{2d} n)$ . Applied separately to the samples  $X$  and  $Y$ , and combined via the triangle inequality, this yields a version of Theorem 1. The proof in [12] relies on the orthogonality of Haar wavelets and is thus limited to the uniform measure on the cube. Our approach avoids orthogonality and applies to arbitrary Borel distributions on  $\mathbb{R}^d$ .

For results on general measures, we refer to [1, 2], which construct low-discrepancy sequences without assumptions on the underlying distribution. However, these constructions do not rely on thinning random samples and are thus not directly applicable to our setting.

Our work also relates to recent advances in online discrepancy minimization [3, 6, 5, 4], which consider assigning a stream of  $2n$  points in  $[0, 1]^d$  to two sets  $X$  and  $Y$  (without discards) to minimize discrepancy (1.1). A variant of our algorithm may offer another approach to that problem. Conversely, it would be interesting to see if their techniques could sharpen Theorem 1.

*Remark 3* (Open question). It remains open whether the bound  $O(\log^{2d} n)$  in Theorem 1 is optimal. It would be interesting to either improve this bound or establish a matching lower bound.

**2. Reduction to the case of uniform marginals.** It suffices to prove Proposition 1 for distributions whose  $d$  coordinate marginals are uniform on  $[0, 1]$ . We now explain why this entails no loss of generality.

Let  $F_1, \dots, F_d$  be the marginal cumulative distribution functions. Without loss of generality, we may assume that each  $F_i$  is strictly increasing. This follows by perturbation: consider a mixture of the original distribution with a Gaussian on  $\mathbb{R}^d$ . More precisely, flip a biased coin with success probability  $\theta$ ; if success, sample from the original distribution; otherwise, sample from the Gaussian. Samples of  $m$  points from the original and perturbed distributions coincide with probability  $\theta^m$ , and so do their sign discrepancies. As  $\theta \rightarrow 1$ , the expected discrepancy of the perturbed sample converges to that of the original. We may thus sample from the mixture instead of the original distribution, and the marginals  $F_i$  of the mixture are strictly increasing.

For a real-valued random variable  $X$  with cumulative distribution function  $F$ , define the (randomized) integral transform of  $X$  as a random convex combination of  $F(X)$  and  $F(X^-)$ :

$$\hat{X} := UF(X) + (1 - U)F(X^-),$$

where  $U$  is an independent random variable uniformly distributed on  $[0, 1]$ , and  $F(x^-) = \lim_{y \uparrow x} F(y)$ . Then  $\hat{X}$  is uniformly distributed on  $[0, 1]$ . Applying this transform coordinate-wise to any probability distribution on  $\mathbb{R}^d$  yields one with uniform marginals.

Let  $x_1, \dots, x_m$  be an arbitrary set of points in  $\mathbb{R}^d$ . For each  $x_i \in [0, 1]^d$ , consider its transformed version:

$$\hat{x}_i(k) = U_i F_k(x_i(k)) + (1 - U_i)F_k(x_i(k)^-),$$

where  $x_i(k)$  denotes the  $k$ th coordinate of  $x_i$ . We use its own independent copy  $U_i$  for each point. This preserves independence when the  $x_i$  are sampled independently from the original distribution.

**Lemma 1.** *If each  $F_k$  is strictly increasing, then the sign discrepancy of  $\hat{x}_1, \dots, \hat{x}_m$  is no less than that of  $x_1, \dots, x_m$ .*

*Proof.* To each axis-aligned half-infinite box  $B = (-\infty, b_1] \times \dots \times (-\infty, b_d]$ , associate the anchored box  $\hat{B} = [0, F_1(b_1)] \times \dots \times [0, F_d(b_d)]$ . It suffices to show that for each  $i$ ,

$$x_i \in B \iff \hat{x}_i \in \hat{B}.$$

Indeed, if  $x_i \in B$ , then  $\hat{x}_i \in \hat{B}$ , regardless of  $U_i$ . If  $x_i \notin B$ , then  $x_i(k) > b_k$  for some  $k$ , hence  $F_k(x_i(k)) > F_k(b_k)$  by strict monotonicity. Thus,  $\hat{x}_i(k) \geq F_k(x_i(k)) > F_k(b_k)$ , so  $\hat{x}_i \notin \hat{B}$ . The completes the proof.  $\square$

We conclude:

**Lemma 2.** *If Proposition 1 holds for all Borel distributions with uniform marginals, then it holds for all Borel distributions.*

*Proof.* Using the perturbation argument above, we reduce to the case where all marginals have strictly increasing distribution functions. Apply the randomized transform to obtain  $\hat{x}_1, \dots, \hat{x}_m$  from  $x_1, \dots, x_m$ , using independent  $U_i$  to preserve independence.

Let us discard some of the transformed points  $\hat{x}_i$  to reduce discrepancy. Then discard the corresponding indices from the original sample  $x_i$ . By Lemma 1, which does not require independence, the discrepancy of the retained  $x_i$  is no greater than that of the retained  $\hat{x}_i$ . This completes the proof.  $\square$

**3. Proof of Proposition 1.** Our proof relies on the following statement, which may be of independent interest.

**Proposition 2.** *Consider any vectors  $v_1, \dots, v_m \in \mathbb{R}^{\dim}$  and a number  $T > 0$ . Let  $\varepsilon_1, \dots, \varepsilon_m$  be independent Rademacher random variables. There is an online<sup>2</sup> algorithm to discard terms  $\varepsilon_i v_i$  so that the running sums of the remaining terms  $\varepsilon'_i v'_i$  satisfy*

$$\sup_k \left\| \sum_{i=1}^k \varepsilon'_i v'_i \right\|_{\infty} \leq T \quad \text{deterministically,}$$

and

$$\mathbb{E} \# \{ \text{discarded terms} \} \leq \frac{1}{T} \sum_{i=1}^m \|v_i\|_1.$$

The proof will need an elementary lemma:

**Lemma 3.** *Fix any vector  $v \in \mathbb{R}^{\dim}$  and a number  $T > 0$ . Let  $w$  be a random vector uniformly distributed in the cube  $[-T/2, T/2]^{\dim}$ . Then*

$$\mathbb{P} \{ \|w + v\|_{\infty} > T/2 \} \leq \frac{\|v\|_1}{T}.$$

*Proof.* By union bound, we have

$$\mathbb{P} \{ \|w + v\|_{\infty} > T/2 \} \leq \sum_{i=1}^{\dim} \mathbb{P} \{ |w_i + v_i| > T/2 \}. \quad (3.1)$$

To compute the  $i$ -th term in this sum, we can assume by symmetry that  $v_i \geq 0$ . Since  $w_i$  is uniformly distributed in  $[-T/2, T/2]$ , we have

$$\mathbb{P} \{ |w_i + v_i| > T/2 \} = \mathbb{P} \{ w_i > T/2 - v_i \} = \frac{v_i}{T}.$$

Thus, each term in the sum in (3.1) is bounded by  $|v_i|/T$ . This completes the proof.  $\square$

*Proof of Proposition 2.* Let  $K := [-T/2, T/2]^{\dim}$ . We arrange the discards as follows:

- Sample  $v_0 \in K$  uniformly at random, and set  $w_0 = v_0$ .
- Upon receiving  $\varepsilon_k v_k$ , check whether

$$w_{k-1} + \varepsilon_k v_k \in K.$$

If so, accept the term  $\varepsilon_k v_k$  and update  $w_k \leftarrow w_{k-1} + \varepsilon_k v_k$ ; otherwise, discard the term and set  $w_k = w_{k-1}$ .

- Repeat for  $\varepsilon_k v_{k+1}$ .

---

<sup>2</sup>The ‘‘online’’ means that, at time  $i$ , the algorithm decides whether to discard the term  $\varepsilon_i v_i$  based on this and previously retained terms; the decision is never reversed in the future.

This algorithm outputs a thinned subsequence of the original stream. By design, all  $w_k$  remain in  $K$  at all times  $k = 0, 1, 2, \dots$ , so all running sums of the accepted terms satisfy  $\|w_k - w_0\|_\infty \leq \|w_k\|_\infty + \|w_0\|_\infty \leq T/2 + T/2 \leq T$ , as claimed.

The random walk  $w_0, \dots, w_m$  is a Markov chain on  $K$ . Due to the symmetry of the Rademacher distribution, the Markov chain is symmetric:

$$\mathbb{P}\{w_k = w' \mid w_{k-1} = w\} = \mathbb{P}\{w_k = w \mid w_{k-1} = w'\} \quad \text{for each } w, w' \in K.$$

Consequently, the uniform distribution on  $K$  is stationary. Since  $w_0$  is initialized uniformly, each  $w_k$  remains uniformly distributed on  $K$ . By design, the term  $\varepsilon_k v_k$  is discarded if and only if  $\|w_{k-1} + \varepsilon_k v_k\|_\infty > T/2$ . Thus:

$$\mathbb{E} \# \{\text{discarded terms}\} = \sum_{k=1}^m \mathbb{P}\{\|w_{k-1} + \varepsilon_k v_k\|_\infty > T/2\} \leq \frac{1}{T} \sum_{k=1}^m \|v_k\|_1,$$

where the last bound follows from Lemma 3, since  $w_{k-1}$  is uniformly distributed in  $K$  and is independent of  $\varepsilon_k$ . The proof is complete.  $\square$

*Proof of Proposition 1.* We may assume that the distribution is supported on  $[0, 1]^d$  and has uniform marginals – see §2. The plan is to first bound the discrepancy over dyadic boxes using Proposition 2, then extend the bound to lattice boxes via a union bound, and finally generalize it to all boxes through approximation.

Fix a resolution parameter  $L \geq 1$ ; we will later choose it as  $L = \log_2 n$  but it is simpler to keep it general for now.

*Step 1. Dyadic boxes.* A dyadic interval is an interval of the type

$$\left[ \frac{j}{2^\ell}, \frac{j+1}{2^\ell} \right] \quad \text{where } j \in \{0, 1, \dots, 2^\ell - 1\} \text{ and } \ell \in \{0, 1, \dots, L\}.$$

A dyadic box is the product of  $d$  dyadic intervals, possibly with different  $j$  and  $\ell$  in different factors. To each point  $x_i \in [0, 1]^d$  of our sample, associate the vector that encodes which dyadic boxes the point belongs to:

$$v_i := (\mathbf{1}_{\{x_i \in B\}})_{B: \text{dyadic box}} \in \mathbb{R}^{\dim}, \quad \dim = (2^{L+1} - 1)^d.$$

This vector is indexed by all dyadic boxes  $B$ , with each coordinate equal to 1 if  $x_i$  lies in  $B$  and 0 otherwise. It follows from the uniform marginal assumption that a.s. no  $x_i$  lies on the boundary of a dyadic box. Then:

$$\|v_i\|_1 = \# \{\text{dyadic boxes } B \text{ that contain } x_i\} = L^d.$$

Apply Proposition 2 with  $TL^d$  instead of  $T$ . With

$$\mathbb{E} \# \{\text{discarded terms}\} \leq \frac{1}{TL^d} \sum_{i=1}^m \|v_i\|_1 \leq \frac{mL^d}{TL^d} \leq \frac{m}{T},$$

as required, the remaining terms  $\varepsilon'_i v'_i$  satisfy

$$\sup_k \left\| \sum_{i=1}^k \varepsilon'_i v'_i \right\|_\infty \leq TL^d \quad \text{deterministically.}$$

By definition of  $v_i$  above and sign discrepancy in (1.2), this means that the remaining points of the sample satisfy

$$\text{disc}_k(B) \leq TL^d \quad \text{for each dyadic box } B \text{ and each } k. \quad (3.2)$$

*Step 2. Lattice boxes.* A lattice interval is an interval of the type

$$\left[0, \frac{j+1}{2^L}\right] \quad \text{where } j \in \{0, 1, \dots, 2^L - 1\}.$$

A lattice box is the product of  $d$  lattice intervals, possibly with different  $j$  in different factors.

Any lattice interval can be partitioned into at most  $L$  dyadic intervals. Thus, any lattice box can be partitioned into at most  $L^d$  dyadic boxes. Therefore, summing  $L^d$  bounds (3.2) by triangle inequality, we conclude:

$$\text{disc}_k(B) \leq TL^{2d} \quad \text{for each lattice box } B \text{ and each } k. \quad (3.3)$$

*Step 3. Slices.* We are about to approximate a general box by a dyadic box, with the approximation error controlled by the number of points that can fall into a thin slice. To prepare for this, let's bound that number.

A slice is the product of one interval of the type

$$\left[\frac{j}{2^L}, \frac{j+1}{2^L}\right], \quad \text{where } j \in \{0, 1, \dots, 2^L - 1\}, \quad (3.4)$$

and  $d-1$  full intervals  $[0, 1]$ ; the interval can appear as any of the  $d$  factors.

By the assumption made at the beginning of the proof, the distribution of the points  $x_i$  has uniform marginals in  $[0, 1]$ . Thus, for any slice  $S$ , the number of points falling in  $S$ ,

$$Z_S := \#\{i \in [m] : x_i \in S\},$$

has binomial distribution with parameters  $m$  and  $2^{-L}$ . Since increasing  $m$  can only increase the number of points in slices, we assume the worst case  $m = 2n$ . Applying Chernoff's inequality [15, Theorem 2.3.1], we get for each  $u > 0$ :

$$\mathbb{P}\{Z_S > 16n2^{-L} + \ln(d2^L) + u\} \leq e^{-\ln(d2^L) - u} = \frac{e^{-u}}{d2^L}.$$

Taking the union bound over all  $d2^L$  possible slices, we get

$$\mathbb{P}\left\{\max_S Z_S > 16n2^{-L} + \ln(d2^L) + u\right\} \leq e^{-u}.$$

In other words, with probability at least  $1 - e^{-u}$ , the following event holds:

$$\max_{S: \text{slice}} \#\{i \in [2n] : x_i \in S\} \leq 16n2^{-L} + \ln(d2^L) + u. \quad (3.5)$$

*Step 4. General boxes.* Any interval  $B = [0, b] \subset [0, 1]$  can be approximated by some lattice interval  $B' \supset B$  by rounding  $b$  up to a lattice point; this way  $B' \setminus B$  lies in some dyadic interval of the type (3.4). Therefore, a general box

$$B = [0, b_1] \times \dots \times [0, b_d]$$

can be approximated by a lattice box  $B' \supset B$  so that  $B' \setminus B$  lies in a union of  $d$  slices. By triangle inequality, it follows that

$$\text{disc}_k(B) \leq \text{disc}_k(B') + d \cdot \max_{S: \text{slice}} \#\{i \in [2n] : x_i \in S\}.$$

Taking supremum over axis-aligned boxes  $B$  and  $k$ , and using (3.3), we find:

$$\sup_{B: \text{box}, k} \text{disc}_k(B) \leq TL^{2d} + d \cdot \max_{S: \text{slice}} \#\{i \in [2n] : x_i \in S\}.$$

Set  $L = \log_2 n$ . From (3.5), with probability  $1 - e^{-u}$ , we find:

$$\sup_{B: \text{box}, k} \text{disc}_k(B) \leq T \log_2^{2d} n + d \ln(dn) + d(u + 16).$$

Integrating the tail (see [15, Lemma 1.2.1]) completes the proof.  $\square$

## References.

- [1] C. Aistleitner, D. Bilyk, and A. Nikolov. Tusnády’s problem, the transference principle, and non-uniform QMC-sampling. *CoRR* 17.
- [2] C. Aistleitner and J. Dick. Low-discrepancy point sets for non-uniform measures. *Acta Arithmetica*, 163(4):345–369, 2014.
- [3] R. Alweiss, Y. Liu, and M. Sawhney. Discrepancy minimization via a self-balancing walk. STOC 21, page 14–20. ACM.
- [4] N. Bansal, H. Jiang, R. Meka, S. Singla, and M. Sinha. Online discrepancy minimization for stochastic arrivals. In *Proc. SODA 2021*, pages 2842–2861, 2021.
- [5] N. Bansal, H. Jiang, S. Singla, and M. Sinha. Online vector balancing and geometric discrepancy. STOC 20, page 1139–1152. ACM.
- [6] N. Bansal and J. Spencer. On-line balancing of random inputs. *Rand. Struct. Alg.*, 57(4):879–891, 2020.
- [7] J. Beck and W. Chen. *Irregularities of Distribution*. Cambridge University Press, 1987.
- [8] D. Bilyk, M. Mastrianni, R. Matzke, and S. Steinerberger. Polarization and greedy energy on the sphere, 2023.
- [9] D. Bilyk and S. Steinerberger. Regularizing random points by deleting a few, 2025.
- [10] B. Chazelle. *The Discrepancy Method*. Cambridge University Press, 2000.
- [11] F. Clement, C. Doerr, and L. Paquete. Heuristic approaches to obtain low-discrepancy point sets via subset selection. *Journal of Complexity*, 83, 2024.
- [12] R. Dwivedi, O. Feldheim, O. Gurel-Gurevich, and A. Ramdas. The power of online thinning in reducing discrepancy. *Prob. Th. Rel. Fields*, 174(1–2):103–131, 2018.
- [13] J. Matoušek. *Geometric Discrepancy*, volume 18 of *Algorithms and Combinatorics*. Springer-Verlag, 1999.
- [14] G. Smirnov and R. Vershynin. Improving discrepancy by moving a few points, 2025.
- [15] R. Vershynin. *High-dimensional Probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

MATHEMATICAL SCIENCES INSTITUTE, AUSTRALIAN NATIONAL UNIVERSITY, CANBERRA, AUSTRALIA  
*Email address:* gleb.smirnov@anu.edu.au

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF CALIFORNIA, IRVINE, US  
*Email address:* rvershyn@uci.edu