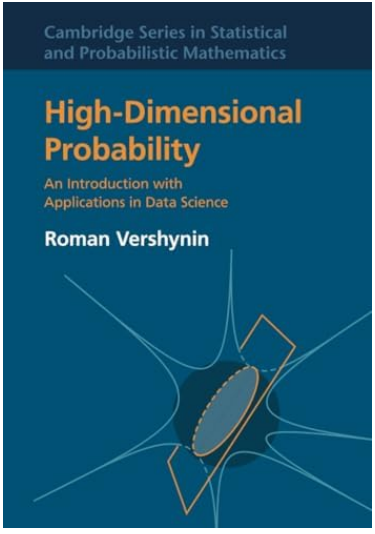
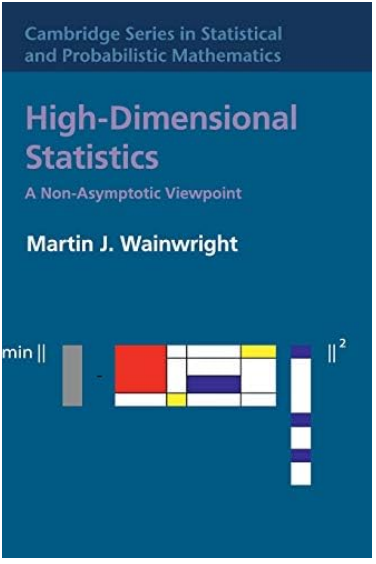
High dimensions: Curse or Blessing for Statisticians?

Luis Sierra Muntané
luis.sierra@mail.utoronto.ca
Department of Statistical Sciences, University of Toronto

Statistical Methodology Reading Group

9/16th October 2026





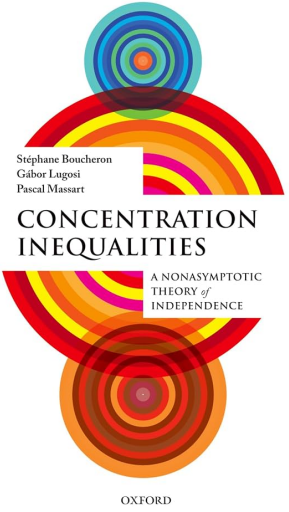
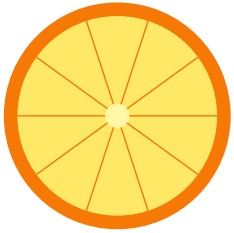


Figure: Three JPG images displayed side by side.

A high-dimensional orange



radius 1, peel 1% thick
(not to scale)

Suppose we have an orange of radius 1 in the space $\mathbb{R}^{1000}$. You peel off a skin 1% thick.

How much of the orange's volume was removed?

A about 1%

B about 10%

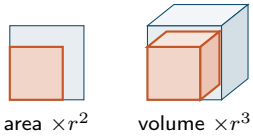
C about half

D over 99%

99.996% of it is peel.

A high-dimensional orange

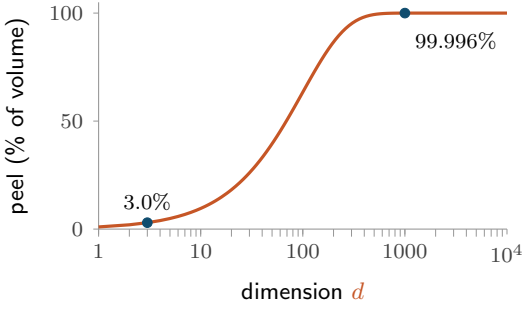
Shrink a shape by a factor r along every axis:



Each of the d axes contributes a factor r, so $\mathrm{vol}(rB) = r^d \,\mathrm{vol}(B)$.

The pulp is a ball of radius 0.99, so

$$\text{peel} = 1 - 0.99^d \approx 1 - e^{-d/100}.$$



> *In high dimensions, almost all of the volume of a ball is close to its surface.*

Curse or blessing?

The term *curse of dimensionality* is at least as old as Bellman (1957). Donoho (2000) however, notes that high dimensionality also brings with it some *blessings*.

	Curse	**Mechanism**	**Blessing**
Empty space	?	?	?
Nearby distributions	?	?	?

We start with two problems that get worse as the dimension grows, and look for the reason. The same reason will then give us two methods that make use of, and improve with dimensionality.

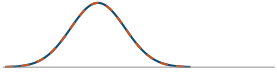
Too many directions and k-nearest neighbours

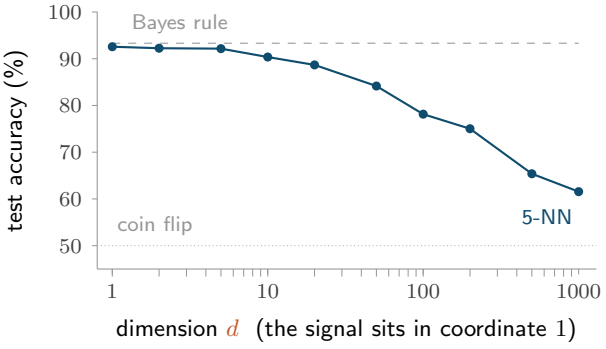
Consider two classes $\mathcal{N}(0,1)$ and $\mathcal{N}(3,1)$ with 100 observations in each.

coordinate 1: signal



coordinates $2,\ldots,d$: noise



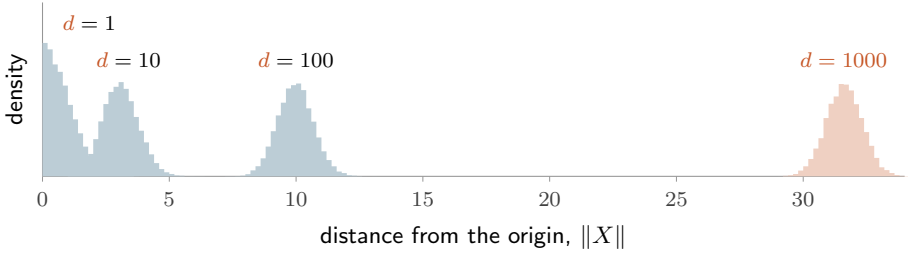


Why. $\|x - x'\|^2 = (x_1 - x_1')^2 + \sum_{i\geq 2}(x_i - x_i')^2$. Between classes the first term is larger by about $3^2 = 9$. Each noise term has mean 2 and variance 8, so their sum is $2(d-1) \pm \sqrt{8(d-1)}$. Once $\sqrt{8d} \gg 9$, the noise has a larger influence on the distance.

> *Smoothing-like procedures have a similar problem: a kernel density estimate at the mode of $\mathcal{N}(0, I_{10})$ needs about $842{,}000$ points to bring its relative mean squared error below 0.1 (Silverman, 1986).*

Where is a Gaussian distribution concentrated?

$X \sim \mathcal{N}(0, I_{1000})$. Its density is highest at the origin. **How close to the origin do you expect to see a random draw?**



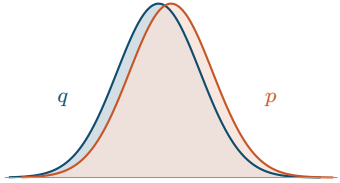
The histograms are at $\sqrt{d}$ because $\mathbb{E}\|X\|^2 = \sum_{i=1}^{d} \mathbb{E}\, X_i^2 = d$, and their widths are bounded! ($\mathrm{Var}(\|X\|) = d - (\mathbb{E}\|X\|)^2 = d - (\sqrt{d} - 1/\sqrt{d})$). At $d = 1000$, 98% of draws land in $[30.0, 33.3]$, and $\mathbb{P}(\|X\| \leq 10) \approx 10^{-306}$.

> *Almost no draws are found near the mode but rather lie on a thin shell, the Gaussian soap bubble.*

Experiment with Importance Sampling

Suppose we want to calculate $\mathbb{E}_p f(X)$ using samples from q: (abusing notation of distribution and density)

$$\mathbb{E}_p f(X) = \int f\, p = \int f \frac{p}{q}\, q = \mathbb{E}_q \left[f(X)\, w(X) \right].$$



Draw $X_1, \ldots, X_N \sim q$ and take the weighted average of $f(X_i)$ with weights $w(X_i) = \frac{p}{q}(X_i)$.
Simulation: $q = \mathcal{N}(0, I_d)$ and $p = \mathcal{N}(0.3 \cdot \mathbf{1}, I_d)$. In any coordinate, the probability of overlap is 88%.

Take $N = 10{,}000$ draws from q and calculate the weights $w(x_i)$. If $d = 1000$, how many weights do we need to add up to reach 50% of the total proportion?

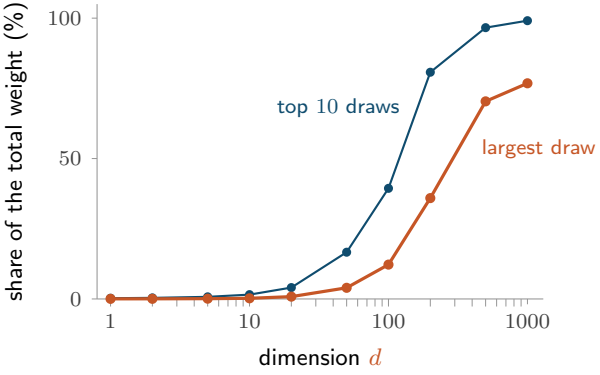
A more than $1/2$ (5000)

B about 1000

C about 100

D about one

Distribution of the weights



Out of $10{,}000$ draws, the top 10 carry 2% of the weight at $d = 10$ but 39% at $d = 100$.
At $d = 1000$ a single draw carried 77% of the weight.
The estimate of $\mathbb{E}_p X_1 = 0.3$ is off by 0.74, which is worse than answering 0.

Why?
Note that for our two Gaussians,

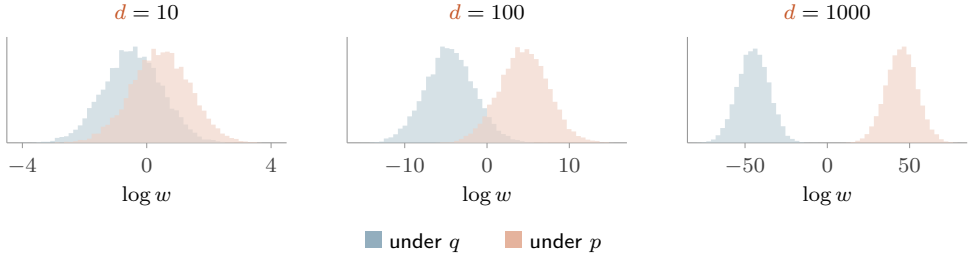
$$w(x) = \frac{p(x)}{q(x)} = \frac{e^{-\|x-\mu\mathbf{1}\|^2/2}}{e^{-\|x\|^2/2}} = \exp\Big(\sum_{i=1}^{d}\big(\mu x_i - \tfrac{\mu^2}{2}\big)\Big).$$

with $\mu = 0.3$.

Intuitive explanation for the weight distribution

This means $\log w$ is the sum of d independent terms with variance μ^2 and mean $\mp\frac{\mu^2}{2}$ (under q / under p):

$$\begin{cases} \log w(x) \sim \mathcal{N}\left(-\frac{d\mu^2}{2}, d\mu^2\right), & \text{when } x \sim q, \\ \log w(x) \sim \mathcal{N}\left(+\frac{d\mu^2}{2}, d\mu^2\right), & \text{when } x \sim p. \end{cases} \quad \Longrightarrow \quad \log w \approx \mp\frac{d\mu^2}{2} \pm \mu\sqrt{d}.$$



The separation between the means scales with the dimension d, while the spreads scale with $\sqrt{d}$. Consequently, draws from q rarely land where the density p is large.

How many draws would we need?

An unweighted average of N many $\mathcal{N}(0,1)$ samples has variance $1/N$. We instead have weights $\{w_i\}$, which after normalizing give a variance

$$\frac{\mathbb{E}_q w^2}{(\mathbb{E}_q w)^2} =: \frac{1}{N_{\text{eff}}}.$$

We had $w(x) = \exp\left(\sum_i (\mu x_i - \frac{\mu^2}{2})\right)$, and under q the x_i are independent $\mathcal{N}(0,1)$. Hance $\mathbb{E}_q\, w = 1$, and the Gaussian moment generating function $\mathbb{E}\, e^{tx_i} = e^{t^2/2}$ gives

$$\mathbb{E}_q w^2 = \mathbb{E}_q \exp\left(2\mu \sum_{i=1}^{d} x_i - d\mu^2\right) = e^{-d\mu^2} \prod_{i=1}^{d} \mathbb{E} e^{2\mu X_i} = e^{-d\mu^2}\, e^{2d\mu^2} = e^{d\mu^2}.$$

The effective sample size therefore scales as $N/N_{\text{eff}} = N/\mathbb{E}_q w^2 = N e^{-d\mu^2}$, so we need $N \gtrsim e^{d\mu^2}$, which is exponential in d.

Theorem (Chatterjee and Diaconis, 2018)

About $e^{\mathrm{KL}(p \,\|\, q)}$ draws are necessary and sufficient (when $\log w$ concentrates). Here $\mathrm{KL}(p \,\|\, q) = \mathbb{E}_p \log w = d\mu^2/2$, so at $d = 1000$ we need $N \approx e^{45} \approx 3.5 \times 10^{19}$.

Common features to our examples

Every quantity we computed is a sum of d independent components:

$$\|X\|^2 = \sum_{i=1}^{d} X_i^2, \qquad \|x - x'\|^2 = \sum_{i=1}^{d} (x_i - x_i')^2, \qquad \log w = \sum_{i=1}^{d} \left(\mu X_i - \tfrac{\mu^2}{2}\right).$$

Means add and variances add, so a sum of d components is of order $O(d)$ with fluctuations $O(\sqrt{d})$.

sum of d pieces	mean	sd	so we predict	simulated
$\|X\|^2 = \sum X_i^2$	d	$\sqrt{2d}$	$\|X\| \approx \sqrt{d} \pm 0.71$	± 0.71 ✓
$\langle X, Y\rangle = \sum X_i Y_i$	0	$\sqrt{d}$	angle $\approx 90^\circ \pm 57.3^\circ/\sqrt{d}$	$\pm 1.8^\circ$ at $d = 1000$ ✓
$\log w = \sum(\mu X_i - \frac{\mu^2}{2})$	$\mp\frac{d\mu^2}{2}$	$\mu\sqrt{d}$	$\mp 4.5 \pm 3$ at $d = 100$	histograms ✓

Row 1: $\|X\| = \sqrt{d + \sqrt{2d}\,Z} \approx \sqrt{d} + Z/\sqrt{2}$ by a Taylor expansion of the square root. Row 2: $\cos(\text{angle}) = \langle X, Y\rangle / (\|X\|\|Y\|) \approx \langle X, Y\rangle / d$, and 1 radian $\approx 57.3^\circ$.

Recap

The more general, informal statement about this concentration phenomenon can be explained in terms of

$$f(X_1, \ldots, X_d) - \mathbb{E}f(X_1, \ldots, X_d)$$

which will be *small* whenever

- The function f is not too *sensitive* on any of its inputs
- The dependence between $X_1, \ldots, X_d$ is not *too strong*.

Functions of Gaussian variables

Observation: $\|X\|$ is concentrated with Gaussian tails. Which property can we use?

$$|\|x\| - \|y\|| \le \|x - y\| :$$

moving x a little moves $\|x\|$ proportionally, so no single coordinate can move it far.

Theorem (Gaussian concentration; Tsirelson et al., 1976)

Let $X \sim \mathcal{N}(0, I_d)$ and let f be L-Lipschitz, that is $|f(x) - f(y)| \le L\|x - y\|$. Then for every $t > 0$,

$$\mathbb{P}\big(|f(X) - \mathbb{E}f(X)| \ge t\big) \;\le\; 2\,e^{-t^2/(2L^2)}.$$

Dimensionless $\Longleftrightarrow$ There is no d on the right-hand side.

Note that if $f(x) = x_1$ then we recover the exponent of a Gaussian tail, so the inequality is tight. For $f(x) = \|x\|$ the bound says that the fluctuations are $O(1)$ irrespective of d, just as we had for the soap-bubble example.

A five-line proof is in the backup slides; see also Boucheron et al. (2013, Thm 5.6).

Can we preserve distances with low-dimensional projections?

Suppose we have $n = 200$ points in $\mathbb{R}^{50\,000}$. Can we map them into $\mathbb{R}^k$ with $k \ll 50\,000$ keeping every pairwise distance within a factor $1 \pm \varepsilon$. **How small can k be?**

$$(1-\varepsilon)\|x_i - x_j\| \le \|f(x_i) - f(x_j)\| \le (1+\varepsilon)\|x_i - x_j\|$$

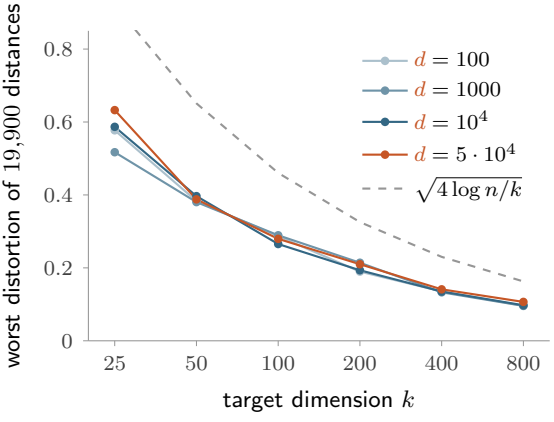
Try a random map: $G \in \mathbb{R}^{k \times d}$ with independent $\mathcal{N}(0, 1/k)$ entries. Its rows are random directions, so they are nearly orthogonal.

1. **Take a pair**, $u = x_i - x_j$: since $Gu \sim \mathcal{N}\left(0, \frac{\|u\|^2}{k} I_k\right)$, $\frac{\|Gu\|}{\|u\|} \overset{d}{=} \frac{\|g\|}{\sqrt{k}}$ with $g \sim \mathcal{N}(0, I_k)$.
2. **Apply the tail bound:** $f = \|\cdot\|$, $L = 1$, $t = \varepsilon\sqrt{k}$, $\mathbb{E}\|g\| \approx \sqrt{k}$: $\mathbb{P}\left(\left|\frac{\|Gu\|}{\|u\|} - 1\right| \ge \varepsilon\right) \lesssim 2e^{-k\varepsilon^2/2}$.
3. **Union bound:** $n^2/2$ pairs, $\mathbb{P}(\text{some pair fails}) \le n^2 e^{-k\varepsilon^2/2} < 1$ once $k > \frac{4 \log n}{\varepsilon^2}$.

Johnson and Lindenstrauss (1984)

$k = O(\log n / \varepsilon^2)$ dimensions suffice, whatever d is. The $\log n$ is the price from the union bound.

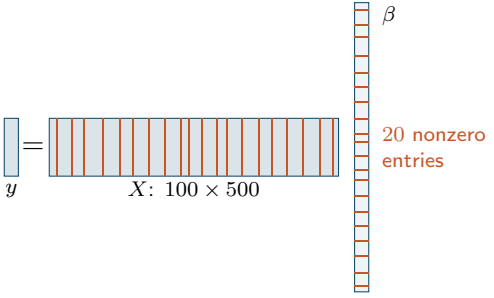
Johnson–Lindenstrauss in practice



- The four curves coincide: the distortion does not depend on d.
- At $k = 200$ the worst of all $19{,}900$ distances is off by 0.19–0.21.
- At $k = 800$, by 0.10–0.11.
- The dashed union-bound curve is cautious by a factor of about 1.6.

Example: sparse linear model with $p > n$

Consider the underdetermined system $y = X\beta$ with $X \in \mathbb{R}^{100\times 500}$. Taking $(x_i^\top)_i$ as rows of X, $y_i = \langle x_i, \beta\rangle$ are measurements of the unknown vector β.

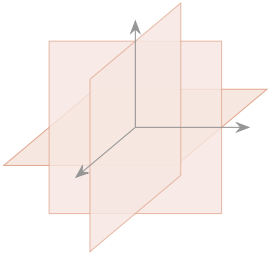


Unless β is sparse: suppose only $s = 20$ of its entries are nonzero.

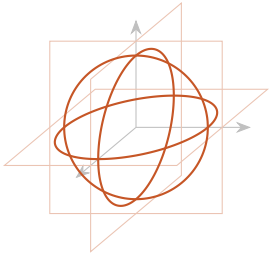
- Then y is a combination of just 20 columns of X.
- If we knew which, we would have 100 equations in 20 unknowns.
- **But we don't know which.**

How can we use dimensionality for sparsity?

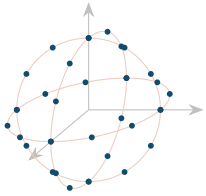
Recall from the JL lemma that $k \approx 2 \log N/\varepsilon^2$ random rows maintain the lengths of N vectors. To determine a sparse β, $X/\sqrt{n}$ must maintain the length of *every* sparse vector. **Is this the case?**



1. All sparse vectors. They fill $\binom{p}{s}$ coordinate subspaces of dimension s. Here $p = 3$, $s = 2$: three planes.



2. Unit vectors suffice. X is linear, so $\|X(cv)\| = |c| \, \|Xv\|$. Keep one unit sphere per subspace.



3. Extract a finite net. $(3/\varepsilon)^s$ points ε apart can be packed on each sphere. X distorts norms by less than 3ε.

$\implies$ **Apply JL to** $h \in \mathcal{P}\left(\bigcup_{|T|=s-1} \mathbb{S}_T\right)$**:** at most $\binom{p}{s}(3/\varepsilon)^s$ "objects" which gives us $\log\left[\binom{p}{s}(3/\varepsilon)^s\right] \leq s \log \frac{ep}{s} + s \log \frac{3}{\varepsilon} \implies n \gtrsim s \log \frac{ep}{s}$ random equations keep the length of every sparse vector at once (Candès et al., 2006; Donoho, 2006; Baraniuk et al., 2008).

> *Here* $\binom{500}{20} \approx 2.7 \times 10^{35}$ *subspaces, but* $s \log \frac{ep}{s} \approx 84$.

Finding the sparse solution: why ℓ_1?

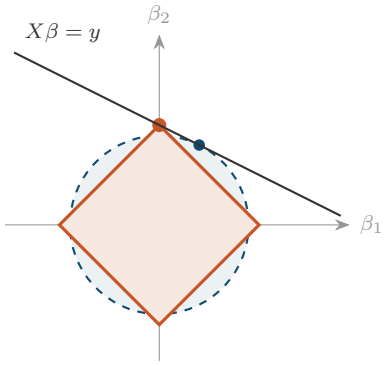
Trying all 2.7×10^{35} supports by solving 10^9 problems per second would take $8.5 \times 10^{1}8$ years $\implies$ hopeless.
We can instead use the convex relaxation:

$$\hat{\beta} = \arg\min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad X\beta = y,$$

which is a linear and **convex** program.

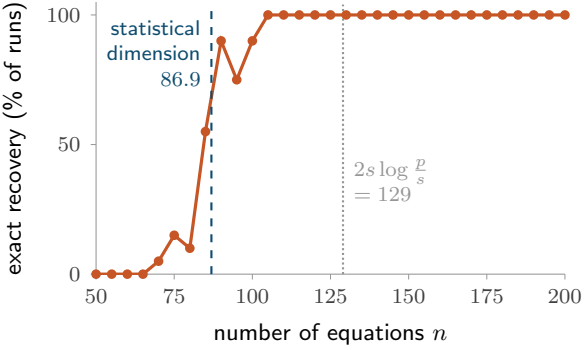
Algorithm idea. Inflate the ℓ_1 ball until it intersects the solutions: it touches at a corner, and the corners are the sparse vectors.

The round ℓ_2 ball touches at a generic point, which is not sparse.



How many equations does ℓ_1 need?

$p = 500$ unknowns, $s = 20$ of them nonzero, Gaussian X. **At what n does ℓ_1 start recovering β reliably?**



500 unknowns from about 90 equations.

The jump is at the *statistical dimension* of the problem (Amelunxen et al., 2014).

Further numerical experiments with a fixed sparsity ratio:

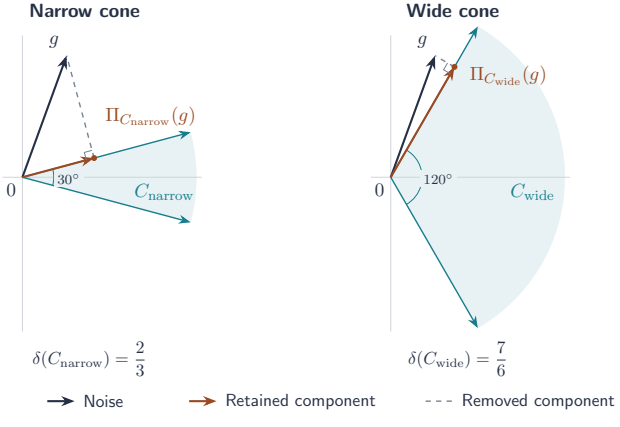
- $s/p = 4\%$ nonzero, Gaussian X, n equations (rows of X).
- Exact recovery (% of runs). The statistical dimension puts the edge at $n/p \approx 0.174$.

Exact recovery (%)

p	s	n/p			
		0.12	0.16	0.20	0.24
25	1	34	57	73	86
100	4	7	42	82	95
500	20	0	21	93	100
2000	80	0	4	100	100

Allowed directions and statistical dimension

The same Gaussian noise vector, projected onto two different cones



Both cones span $\mathbb{R}^2$, but the wider cone retains more noise on average.

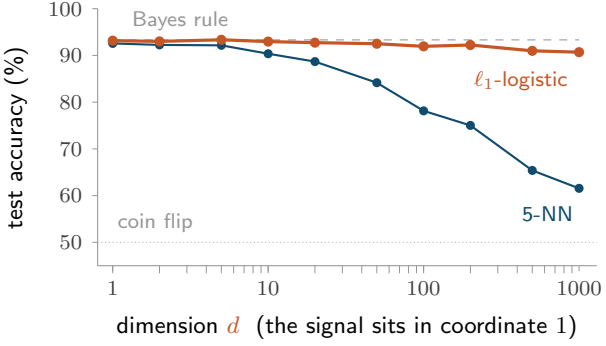
$$\delta(C) = \mathbb{E}_{g\sim N(0,I_2)}\left[\|\Pi_C(g)\|^2\right] \qquad \delta(C_\alpha) = \frac{1}{2} + \frac{\alpha}{\pi} \quad (0 \le \alpha \le \pi)$$

Here α is the full opening angle in radians. Boundary rays extend indefinitely.

Recall our problem of separating two classes in $\mathbb{R}^d$, distributed as $\mathcal{N}(0,1)$ and $\mathcal{N}(3,1)$ on the first coordinate, but $\mathcal{N}(0, I_{d-1})$ on all others, with 100 observations in each.

Back to k-nearest neighbours

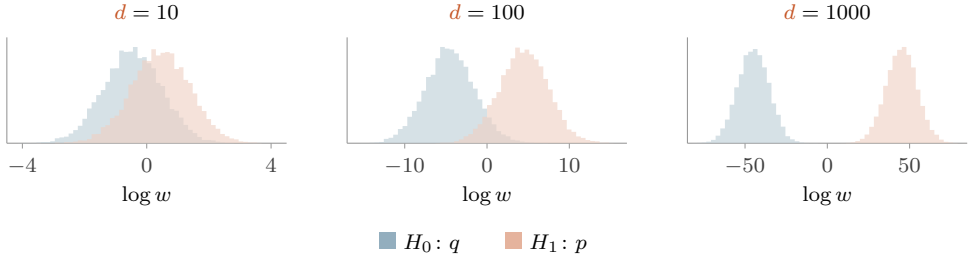
Recall our problem of separating two classes in $\mathbb{R}^d$, distributed as $\mathcal{N}(0, 1)$ and $\mathcal{N}(3, 1)$ on the first coordinate, but $\mathcal{N}(0, I_{d-1})$ on all others, with 100 observations in each.



On the same data, ℓ_1-penalized logistic regression stays at 91–93% while 5-NN falls to 62%. A method that relied on neighbourhoods performed poorly in high-dimensions. A method that is able to adapt to a lower "intrinsic" dimensionality performs much better.

Back to our two Gaussians

In importance sampling, the separation lead to a challenge, but a separation can be exploited in a different setting. The two Gaussians $\mathcal{N}(0, I_d)$, $\mathcal{N}(\mu \cdot \mathbf{1}, I_d)$ differ a little in every coordinate.



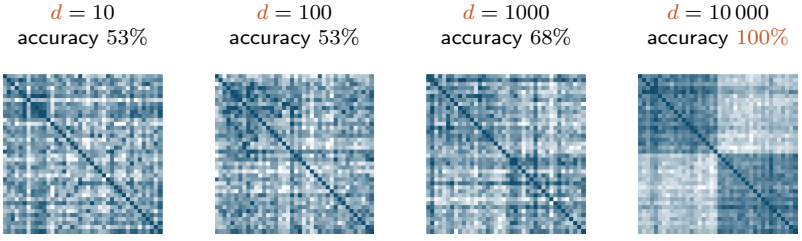
Reject $H_0 : q$ in favour of $H_1 : p$ when $\log w > 0$. Under H_0, $\log w \approx \mathcal{N}\big(-\frac{d\mu^2}{2}, d\mu^2\big)$, so the error is $\Phi\left(-\frac{d\mu^2/2}{\mu\sqrt{d}}\right) = \Phi\left(-\frac{\mu\sqrt{d}}{2}\right)$: 6.7% at $d = 100$, 1.7% at $d = 200$.

The "affinity" multiplies across coordinates ($e^{-\mu^2/8}$ each), so $\int \sqrt{pq} = e^{-d\mu^2/8} \to 0$ and the laws become mutually singular (Kakutani, 1948).

Clustering without labels and infinite variables

Draw 20 points from each Gaussian and forget which is which. Using distances only:

- From the same group, $\|X - X'\|^2 \approx 2d \pm \sqrt{8d}$
- From different groups, $\|X - Y\|^2 \approx 2d + d\mu^2 \pm \sqrt{8d}$.
- They separate once $d\mu^2 \gg \sqrt{8d}$, that is $d \gg 8/\mu^4 \approx 988$ (Hall et al., 2005).



Distance matrices (dark = close), sorted by the hidden label; accuracy of average-linkage clustering.

> *For q to stand in for p, their typical sets must overlap, and they separate as d grows. To tell q from p, that separation is what we need.*

When dimension hurts, and when it helps

	Curse	Mechanism *concentration of measure*	Blessing
Empty space	k-NN and smoothers need loads of data	lengths and angles concentrate	random projections and sparse recovery
Nearby distributions	importance sampling collapses	$\log \frac{p}{q}$: centre $\propto d$, spread $\propto \sqrt{d}$	testing and clustering become easier

- Dimension hurts methods that need things to stay close: neighbours for k-NN, overlap for importance sampling.
- It helps methods that let evidence add up across coordinates, such as random projections and likelihood-ratio tests.
- We pay for the ambient dimension, and we gain when the effective dimension, such as $s \log p$, is small.

High dimensions can be a curse for questions about neighbourhoods but a blessing when dealing with averages.

To read more. Vershynin (2018), *High-Dimensional Probability*; Wainwright (2019), *High-Dimensional Statistics*; Boucheron et al. (2013), *Concentration Inequalities*; and van Handel (2016) *Probability in High Dimension*.

Thank you!

References I

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA*, 3(3):224–294, 2014.

J. Baik, G. Ben Arous, and S. Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. doi: 10.1093/acprof:oso/9780199535255.001.0001.

E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28 (2):1099–1135, 2018.

D. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A*, 367(1906):4273–4293, 2009.

D. L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. Lecture at the AMS conference *Mathematical Challenges of the 21st Century*, 2000.

References II

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

D. L. Donoho, M. Gavish, and I. M. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *The Annals of Statistics*, 46(4):1742–1778, 2018.

P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B*, 67(3):427–444, 2005.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

S. Kakutani. On equivalence of infinite product measures. *Annals of Mathematics*, 49(1):214–224, 1948.

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, 1967.

D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4): 1617–1642, 2007.

G. Pisier. Probabilistic methods in the geometry of Banach spaces. In *Probability and Analysis*, volume 1206 of *Lecture Notes in Mathematics*, pages 167–241. Springer, 1986.

B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, 8(6):1348–1360, 1980.

References III

P. Sur and E. J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

M. Talagrand. A new look at independence. *The Annals of Probability*, 24(1):1–34, 1996.

B. S. Tsirelson, I. A. Ibragimov, and V. N. Sudakov. Norms of Gaussian sample functions. In *Proceedings of the Third Japan–USSR Symposium on Probability Theory*, volume 550 of *Lecture Notes in Mathematics*, pages 20–41. Springer, 1976.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

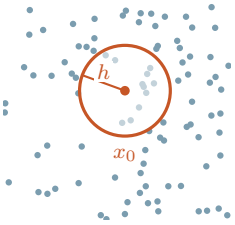
R. van Handel. Probability in high dimension. APC 550 Lecture Notes, Princeton University, 2016. URL `https://web.math.princeton.edu/~rvan/APC550.pdf`. Version dated December 21, 2016.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B*, 76(1):217–242, 2014.

How much data does a smoother need?



count the points in a window

Estimate a density f at x_0 from n points by counting those within distance h:

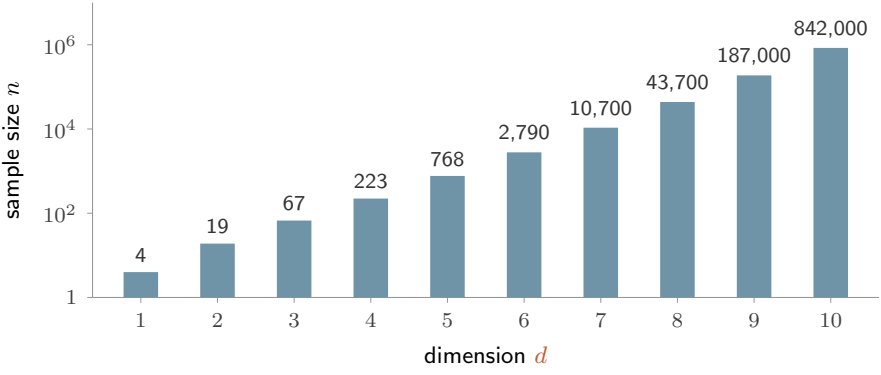
$$\hat{f}(x_0) = \frac{\#\{\text{points in the window}\}}{n \cdot \mathrm{vol}(\text{window})}, \qquad \mathrm{vol}(\text{window}) \propto h^d.$$

- **Bias.** f curves inside the window: the average misses $f(x_0)$ by $\propto h^2$.
- **Variance.** The count has mean $m \approx n f(x_0)\,\mathrm{vol}$ and variance $\approx m$, so $\mathrm{Var}\,\hat{f} \approx \frac{m}{(n\,\mathrm{vol})^2} \propto \frac{1}{nh^d}$.
- **Balance.** Setting $h^4 \asymp \frac{1}{nh^d}$ gives $h \asymp n^{-1/(d+4)}$ and $\mathrm{MSE} \asymp n^{-4/(d+4)}$.

> *To halve the error, multiply n by $2^{(d+4)/4}$: by 11 when $d = 10$, and by 67 million when $d = 100$.*

Exactly how bad? Silverman's table

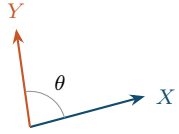
Sample size for a kernel density estimate at the mode of $\mathcal{N}(0, I_d)$ to reach relative MSE 0.1, with the best possible bandwidth:

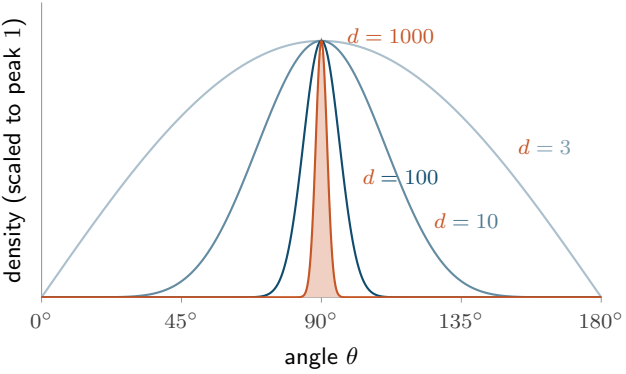


Computed exactly (no asymptotics); the numbers reproduce Silverman (1986, Ch. 4). Stone (1980): no estimator beats the rate $n^{-4/(d+4)}$ for twice-differentiable densities.

Two random directions: what angle?

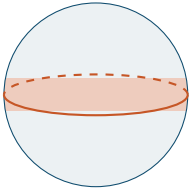
Pick two random directions in $\mathbb{R}^d$.
What is the angle between them?





> *At $d = 1000$ the angle is $90° \pm 1.8°$ (bars: $20{,}000$ simulated pairs). Random directions are almost orthogonal, so there is room for a very large number of them.*

Lévy's picture: most of the sphere is near any equator



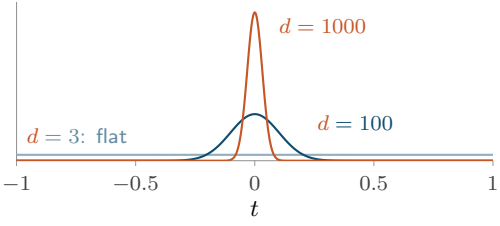
a band of width $\approx 1/\sqrt{d}$ around *any* equator

At $d = 1000$, 99.8% of the sphere lies within 0.1 of the equator.

X is uniform on the unit sphere of $\mathbb{R}^d$. How is one coordinate, $X_1 = t$, distributed?

The slice at height t is a sphere of radius $\sqrt{1-t^2}$ one dimension down, of size $\propto (1-t^2)^{(d-2)/2}$; the slope of the surface adds $1/\sqrt{1-t^2}$:

$$\text{density of } X_1 \;\propto\; (1-t^2)^{(d-3)/2} \;\approx\; e^{-dt^2/2}.$$



> *A quantity built from many independent pieces, none of which matters much on its own, is nearly constant (after Talagrand, 1996).*

Does the theorem make sense?

$$\mathbb{P}\left(|f(X) - \mathbb{E}\, f(X)| \geq t\right) \;\leq\; 2\, e^{-t^2/(2L^2)}$$

Three sanity checks:

1. $f(x) = x_1$. Then $f(X) \sim \mathcal{N}(0,1)$, whose tail is $\approx e^{-t^2/2}$: the bound cannot be improved.
2. $f(x) = (x_1 + \cdots + x_d)/\sqrt{d}$ is 1-Lipschitz by Cauchy–Schwarz: the CLT scale, at every finite d.
3. $f(x) = \|x\|$. Fluctuations of size $O(1)$, whatever d is: ✓ the spread of 0.71 in the norm histograms.

> *Hoeffding and the CLT are about averages and limits. This is about any Lipschitz function, at every finite d: the non-asymptotic viewpoint of Wainwright (2019).*

Backup: Gaussian concentration in five lines

Maurey and Pisier's interpolation (Pisier, 1986), with a worse constant. Let $X, Y \sim \mathcal{N}(0, I_d)$ be independent, f smooth and L-Lipschitz, and for $\theta \in [0, \frac{\pi}{2}]$ put

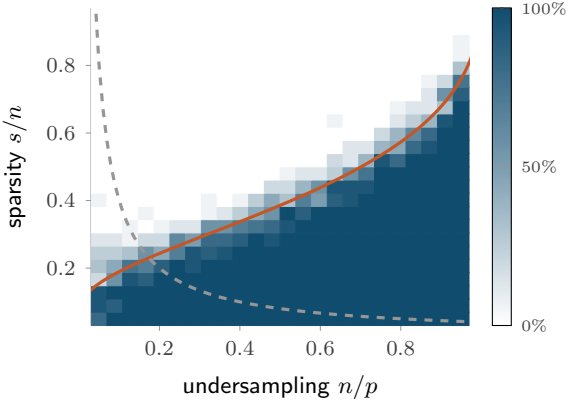
$$X_\theta = X \sin\theta + Y \cos\theta, \qquad X'_\theta = X \cos\theta - Y \sin\theta.$$

1. For each θ, (X_θ, X'_θ) are independent $\mathcal{N}(0, I_d)$: a rotation of (X, Y).
2. $f(X) - f(Y) = \int_0^{\pi/2} \langle \nabla f(X_\theta) \rangle X'_\theta \, d\theta$.
3. Jensen over θ (convexity of $e^{\lambda \cdot}$), then the Gaussian MGF given X_θ:

$$\mathbb{E}\, e^{\lambda(f(X)-f(Y))} \le \frac{2}{\pi} \int_0^{\pi/2} \mathbb{E}\, e^{\lambda \frac{\pi}{2} \langle \nabla f(X_\theta) \rangle X'_\theta} \, d\theta \le e^{\lambda^2 \pi^2 L^2 / 8}.$$

4. Jensen in Y: $\mathbb{E}\, e^{\lambda(f(X) - \mathbb{E} f)} \le \mathbb{E}\, e^{\lambda(f(X)-f(Y))}$.
5. Chernoff with $\lambda = 4t/(\pi^2 L^2)$: $\mathbb{P}\left(f(X) - \mathbb{E} f \ge t\right) \le e^{-2t^2/(\pi^2 L^2)}$. No d anywhere.

The whole phase diagram



Each cell: 16 problems with $p = 200$; dark means ℓ_1 recovered β exactly.
Orange: the theoretical edge, from the statistical dimension (Donoho and Tanner, 2009; Amelunxen et al., 2014).
Dashed: the slice $p = 500$, $s = 20$ from the talk.
The edge is sharp because the statistical dimension itself concentrates.

Backup: one quantity behind the projection results

The *Gaussian width* of a set T of unit vectors, $w(T) = \mathbb{E}\sup_{t\in T}\langle g\rangle t$, measures its effective dimension. A random $G \in \mathbb{R}^{k\times d}$ nearly preserves all of T once

$$k \gtrsim \frac{w(T)^2}{\varepsilon^2} \qquad \text{(Vershynin, 2018, Ch. 7–9).}$$

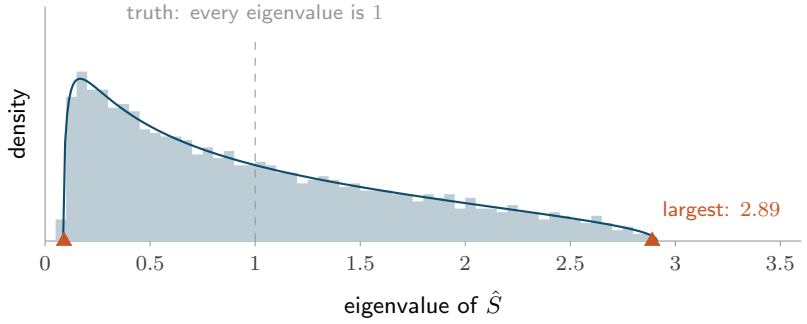
set T	$w(T)^2$	result
n points (their normalised differences)	$\approx 2\log n$	Johnson–Lindenstrauss
s-sparse unit vectors in $\mathbb{R}^p$	$\approx 2s\log(p/s)$	compressed sensing
all unit vectors of $\mathbb{R}^d$	d	no compression possible

The ambient dimension enters only through $w(T)$: effective, not ambient.

What does pure noise look like?

Draw $x_1, \dots, x_n \sim \mathcal{N}(0, I_p)$ with $n = 2000$, $p = 1000$, and form $\hat{S} = \frac{1}{n} \sum_k x_k x_k^\top$. Every true eigenvalue equals 1.

Sketch the histogram of the 1000 eigenvalues of $\hat{S}$.



The sample eigenvalues spread from 0.09 to 2.89: the largest is almost $3\times$ the truth.

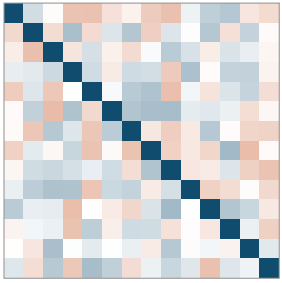
Why so spread? Count the entries

Average eigenvalue: $\frac{1}{p}\sum_j \lambda_j = \frac{1}{p}\operatorname{tr}\hat{S} = \frac{1}{p}\sum_i \hat{S}_{ii} \approx 1.$ ✓
Right on average.

Average squared eigenvalue:
$\frac{1}{p}\sum_j \lambda_j^2 = \frac{1}{p}\operatorname{tr}\hat{S}^2 = \frac{1}{p}\sum_{i,j}\hat{S}_{ij}^2.$

- diagonal: $\hat{S}_{ii} \approx 1$, so it contributes ≈ 1;
- off-diagonal: $\hat{S}_{ij} = \frac{1}{n}\sum_k x_{ki}x_{kj}$ has mean 0 and variance $\frac{1}{n}$, and there are $p(p-1)$ of them: they contribute $\approx \frac{p}{n} = \gamma$.

The eigenvalues have mean 1 and variance $\approx \gamma = \frac{1}{2}$ (simulation: 0.50).



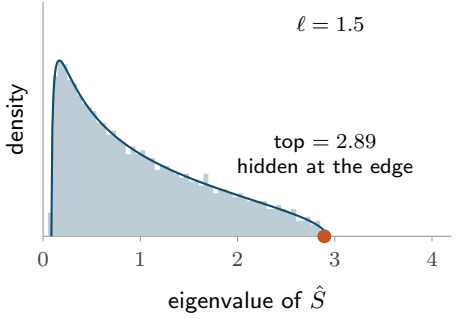
Each off-diagonal entry is tiny, but there are p^2 of them.

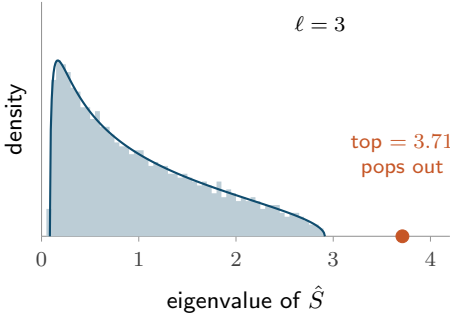
Marčenko and Pastur (1967): the exact shape

Supported on $\left[(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2\right] = [0.09, 2.91]$, so the condition number is ≈ 34 instead of 1: $\hat{S}^{-1}$ is unstable, and so are LDA and Markowitz portfolios (Wainwright, 2019, Ch. 1).

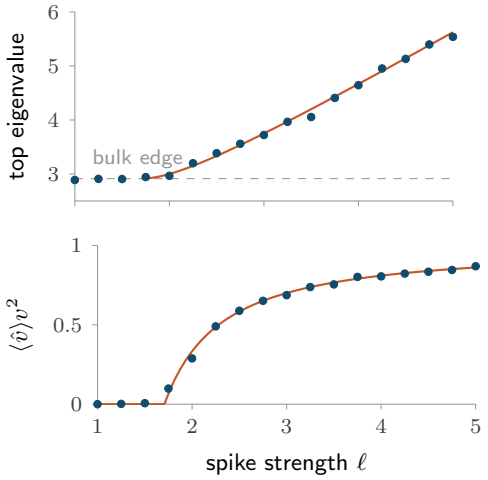
Plant one signal in the noise

$\Sigma = I + (\ell - 1)vv^\top$: one direction with variance ℓ, all others 1. Same $n = 2000$, $p = 1000$.
How large must ℓ be before you can see it?





The BBP transition



Baik et al. (2005) found the curve through the dots:

$$\lambda(\ell) = \ell + \frac{\gamma\,\ell}{\ell - 1}.$$

When does it clear the bulk?
$\lambda'(\ell) = 1 - \frac{\gamma}{(\ell-1)^2}$ vanishes at $\ell^* = 1 + \sqrt{\gamma}$, where $\lambda(\ell^*) = (1 + \sqrt{\gamma})^2$: the bulk edge.
The curve touches the edge at the threshold; below it, the outlier is hidden in the bulk.
Above it, the eigenvector is still off:
$\langle \hat{v} \rangle v^2 \approx 0.70$ at $\ell = 3$, an angle of $33°$ (Paul, 2007).

> *Because the noise bulk is deterministic, we know exactly where noise ends: tests for the number of factors, and optimal eigenvalue shrinkage (Donoho et al., 2018).*

Where this goes next

Benign overfitting

Interpolating the training data with $p > n$ parameters can still generalise, and test error can fall twice as p grows (Bartlett et al., 2020; Hastie et al., 2022).

Classical inference, already at $p/n = 0.2$

The logistic MLE is biased away from zero and the likelihood-ratio test is miscalibrated; both can be corrected (Sur and Candès, 2019).

Also: confidence intervals after the Lasso, via debiasing (Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014).

Backup: the formulas behind the pictures

- **Silverman's table.** Gaussian kernel with bandwidth h, $f = \mathcal{N}(0, I_d)$: $\mathbb{E}\,\hat{f}(0) = (2\pi(1+h^2))^{-d/2}$ and $n \operatorname{Var} \hat{f}(0) = (2\pi h^2)^{-d}(1+2/h^2)^{-d/2} - (\mathbb{E}\,\hat{f}(0))^2$. Minimise the relative MSE over h, then find the smallest n below 0.1.
- **Marchenko–Pastur**, $\gamma = p/n \le 1$, $\lambda_\pm = (1 \pm \sqrt{\gamma})^2$:

$$f_{\mathrm{MP}}(x) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi\gamma x}, \qquad \lambda_- \le x \le \lambda_+.$$

- **BBP** (Baik et al., 2005; Paul, 2007): for $\ell > 1 + \sqrt{\gamma}$, $\lambda_{\max} \to \ell\big(1 + \frac{\gamma}{\ell - 1}\big)$ and $\langle \hat{v} \rangle v^2 \to \dfrac{1 - \gamma/(\ell-1)^2}{1 + \gamma/(\ell-1)}$.
- **Statistical dimension** of the ℓ_1 descent cone at an s-sparse vector (Amelunxen et al., 2014):

$$\delta \approx \inf_{\tau \ge 0} \Big\{ s(1+\tau^2) + (p-s)\,\mathbb{E}\,\big(|g| - \tau\big)_+^2 \Big\}, \qquad g \sim \mathcal{N}(0,1).$$

Backup: questions I expect

Isn't this just the LLN? For sums, yes. Concentration extends it to any Lipschitz function (norms, suprema, eigenvalues, log-likelihoods), with explicit tails at every finite d.

Real data aren't Gaussian. Everything here extends to sub-Gaussian vectors, bounded differences and convex Lipschitz functions of bounded variables (Vershynin, 2018; Boucheron et al., 2013).

Then why does k-NN work on images? Real data often have low *intrinsic* dimension (images lie near a manifold). The curse bites when the intrinsic dimension is high; the blessings need structure such as sparsity or low rank.