Concentration of measure in probability and high-dimensional statistical learning

Lesson # 8

Guillaume Aubrun

Gaussian concentration

Placeholders marked **Proof** will be filled in class by writing on the slides.

Recall from last time

Theorem (Spherical isoperimetric inequality)

Let $A \subset S^{n-1}$ and C be a spherical cap such that $\sigma(A) = \sigma(C)$. Then for every $\varepsilon > 0$, we have $\sigma(A_\varepsilon) \geqslant \sigma(C_\varepsilon)$,

Corollary

Let $f : S^{n-1} \to \mathbf{R}$ a 1-Lipschitz function with median m. Then

$$\sigma(\{|f - m| \geqslant \varepsilon\}) \leqslant \exp(-(n-1)\varepsilon^2/2).$$

Today we are going to prove similar theorems where the sphere S^{n-1} (equipped with the geodesic distance g and the uniform measure σ) is replaced by the Gaussian space, i.e. $\mathbf{R}^n$ equipped with the usual Euclidean distance $|\cdot|$ and the standard Gaussian measure γ_n (= the distribution of $(X_1, \ldots, X_n)$ where X_i are i.i.d. $N(0,1)$).

Fix a dimension n and let $N \geqslant n$. Think of $\mathbf{R}^n$ as a subspace of $\mathbf{R}^N$. Let σ_N be the uniform measure on the sphere $\sqrt{N}S^{N-1}$. Let $\pi_N : \sqrt{N}S^{N-1} \to \mathbf{R}^n$ be the orthogonal projection and μ_N be the image-measure of σ_N under π_N.

Proposition (From the sphere to Gaussian)

The sequence $(\mu_N)_N$ converges in distribution towards γ_n as $N \to \infty$.

Proof

Actually more is true: $\lim_{N\to\infty} \mu_N(B) = \gamma_n(B)$ for every Borel set $(\star)$.

Theorem (Gaussian isoperimetric inequality)

Let $A \subset \mathbf{R}^n$ be a Borel set and H a half-space such that $\gamma_n(A) = \gamma_n(H)$. Then, for every $t > 0$, we have

$$\gamma_n(A_t) \geqslant \gamma_n(H_t).$$

Equivalently, if we define $a \in [-\infty, +\infty]$ by the relation $\gamma_n(A) = \gamma_1((-\infty, a])$, we have $\gamma_n(A_t) \geqslant \gamma_1((-\infty, a+t])$.

Special case : if $\gamma_n(A) = 1/2$ then $a = 0$ and

$$\gamma_n(A_t) \geqslant \gamma_1((-\infty, t])$$

or again

$$\gamma_n(\mathbf{R}^n \setminus A_t) \leqslant \gamma_1([t, +\infty)) = \mathtt{erfc}(t/\sqrt{2}) \leqslant \frac{1}{2}\exp(-t^2/2)$$

If $\gamma_n(A) = 0$ or $\gamma_n(A) = 1$ the result is obvious. Otherwise for every $b < a$, we have $\gamma_n(A) > \gamma_1((\infty, b])$. Consider the projections $\pi_N : \mathbf{R}^N \to \mathbf{R}^n$ and $p_N : \mathbf{R}^N \to \mathbf{R}$. Since

$$\gamma_n(A) = \lim_{N\to\infty} \sigma_N(\pi_N^{-1}(A)) \quad \text{and} \quad \gamma_1((\infty, b]) = \lim_{N\to\infty} \sigma_N(p_N^{-1}((-\infty, b]),$$

we have $\sigma_N(\pi_N^{-1}(A)) \geqslant \sigma_N(p_N^{-1}((-\infty, b]))$ for N large enough.
The spherical isoperimetric inequality implies that

$$\sigma_N(\pi_N^{-1}(A)_t) \geqslant \sigma_N(p_N^{-1}((-\infty, b])_t)$$

where t-enlargements are on $\sqrt{N}S^{N-1}$. We have $\pi_N^{-1}(A)_t \subset \pi_N^{-1}(A_t)$ and

$$p_N^{-1}((-\infty, b])_t) = p_N^{-1}((-\infty, t_N))$$

where t_N is defined by the relations $\sin(\theta_N) = \frac{b}{\sqrt{N}}$ and $\sin(\theta_N + \frac{t}{\sqrt{N}}) = \frac{b+t_N}{\sqrt{N}}$. Since $\lim t_N = t$ (check!), we obtain by $(\star)$

$$\gamma_n(A_t) \geqslant \gamma_1((-\infty, b + t)).$$

The last step is to take the supremum over $b < a$.

As for the sphere, isoperimetry implies concentration for Lipschitz functions

Corollary

Let $f : \mathbf{R}^n \to \mathbf{R}$ be a 1-Lipschitz function with median m with respect to the Gaussian measure γ_n. Then

$$\gamma_n(\{f \geqslant m + t\}) \leqslant \texttt{erfc}(t/\sqrt{2}) \leqslant \frac{1}{2}\exp(-t^2/2).$$

Proof

Equivalently, if $X_1, \ldots, X_n$ are i.i.d. $N(0,1)$ random variables and $Y = f(X_1, \ldots, X_n)$, then $\mathbf{P}(Y \geqslant m_Y + t) \leqslant \frac{1}{2}\exp(-t^2/2)$.

We can replace the median by the expectation.

Corollary

Let $X_1, \ldots, X_n$ are i.i.d. $N(0,1)$ random variables, $f : \mathbf{R}^n \to \mathbf{R}$ a 1-Lipschitz function and $Y = f(X_1, \ldots, X_n)$, then $\mathbf{P}(Y \geqslant \mathbf{E}[Y] + t) \leqslant C \exp(-ct^2)$.

(correct with $C = 1$ and $c = 1/2$)
Example: consider the 1-Lipschitz function $x \mapsto |x|$ on $\mathbf{R}^n$, or $Y = \sqrt{X_1^2 + \cdots + X_n^2}$, so Y^2 has a $\chi^2(n)$ distribution.
We have $\mathbf{E}[Y] \leqslant \mathbf{E}[Y^2]^{1/2} = \sqrt{n}$ and this is sharp (we actually have $\sqrt{n-1} \leqslant m_Y \leqslant \mathbf{E}[Y] \leqslant \sqrt{n}$).
We obtain concentration bounds for χ^2 random variables.

$$\mathbf{P}(Y \geqslant \sqrt{n} + t) \leqslant \frac{1}{2} e^{-t^2/2},$$

$$\mathbf{P}(Y \leqslant \sqrt{n-1} + t) \leqslant \frac{1}{2} e^{-t^2/2}.$$

Such estimates can also be proved by Bernstein inequalities.

High-dimensional data = a finite set $S \subset \mathbf{R}^n$, $n \ggg 1$.

Lemma (Johnson–Lindenstrauss lemma)

Let $S \subset \mathbf{R}^n$ finite, $\varepsilon > 0$. If $k \geqslant 4\varepsilon^{-2} \log \operatorname{card} S$, there is a linear map $f : \mathbf{R}^n \to \mathbf{R}^k$ such that $\forall x, y \in S$,

$$(1-\varepsilon)|x-y| \leqslant |f(x)-f(y)| \leqslant (1+\varepsilon)|x-y|$$

If we are interested in the geometry of S (e.g. we want to identify clusters), we can apply a replace $\mathbf{R}^n$ by $\mathbf{R}^k$ and gain a lot from on computational aspects

Very often $\log \operatorname{card} S \ll n$.

The proof will be by chosing f at random and taking advantage of concentration of measure.

Proof of Johnson–Lindenstrauss lemma

It some situations it is not so obvious to compute either m_Y or $\mathbf{E}[Y]$.
Example: consider a $n \times m$ matrix $M = (Z_{ij})$ with i.i.d. $N(0,1)$ entries, and the function $f : \mathbf{R}^{n \times m} \to \mathbf{R}^+$ mapping M to $\|M\|_{op}$.

$$\|M\|_{op} = \max_{|x|=1} |M(x)| = \max_{|x|=1,|y|=1} \langle Mx, y \rangle.$$

This is a 1-Lipschitz function.
We have $\mathbf{E}\|M\|_{op} \geqslant \max(\sqrt{n-1}, \sqrt{m-1})$.
To show that this is sharp we will rely on comparison theorems for Gaussian processes.

A Gaussian process is a collection $(X_t)_{t\in T}$ of random variables such that any linear combination $\sum \lambda_t X_t$ has a centered Gaussian distribution. Given a Gaussian process $(X_t)_{t\in T}$, the index set T can be equipped with the distance

$$d(s,t) = \left(\mathbf{E}|X_s - X_t|^2\right)^{1/2}$$

Canonical example: if $T \subset \mathbf{R}^n$ and G is a standard Gaussian vector in $\mathbf{R}^n$, one can consider the process $(X_t)_{t\in T}$ defined by $X_t = \langle G, t\rangle$. We have then $d(s,t) = |s-t|$.
Quantity of interest:

$$\mathbf{E} \sup_{t\in T} X_t.$$

Basic example: if $X_1, \ldots, X_n$ are i.i.d. $N(0,1)$ random variables, then

$$\mathbf{E} \sup_{1\leqslant k\leqslant n} X_k = \Theta(\sqrt{\log n})$$

(see Technical Lemma in Lecture 5)

Theorem (Slepian's inequality)

Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be Gaussian processes. Assume that

1. $\mathbf{E}X_t^2 = \mathbf{E}Y_t^2$,
2. $\|X_s - X_t\|_{L^2} \leqslant \|Y_s - Y_t\|_{L^2}$.

Then, for every real numbers (λ_t), we have

$$\mathbf{P}(\exists t \ : \ X_t \geqslant \lambda_t) \leqslant \mathbf{P}(\exists t \ : \ Y_t \geqslant \lambda_t).$$

In particular, $\mathbf{E}\sup_{t\in T} X_t \leqslant \mathbf{E}\sup_{t\in T} Y_t$

The "in particular" part is clear if we know about stochastic domination between random variables X and Y. The following are equivalent

1. $\forall\lambda \in \mathbf{R}, \mathbf{P}(X \geqslant \lambda) \leqslant \mathbf{P}(Y \geqslant \lambda)$,
2. for every increasing function f, $\mathbf{E}f(X) \leqslant \mathbf{E}f(Y)$,
3. there is a coupling (X', Y') such that $\mathbf{P}(X' \leqslant Y') = 1$.

Proof of Slepian's inequality

Proof of Slepian's inequality II

Theorem (Slepian's inequality, second version)

Let $(X_t)_{t \in T}$ and $(Y_t)_{t \in T}$ be Gaussian processes. Assume that

$$\|X_s - X_t\|_{L^2} \leqslant \|Y_s - Y_t\|_{L^2}.$$

Then,

$$\mathbf{E} \sup_{t \in T} X_t \leqslant \mathbf{E} \sup_{t \in T} Y_t$$

Application: norm of Gaussian matrices.
Consider a $n \times m$ matrix $M = (Z_{ij})$ with Z_{ij} i.i.d. $N(0,1)$. We have

$$\mathbf{P}(\|M\|_{op} \geqslant \mathbf{E}[\|M\|_{op}] + t) \leqslant \exp(-t^2/2)$$

with

$$\mathbf{E}\|M\|_{op} = \mathbf{E} \sup_{x \in S^{m-1}, y \in S^{n-1}} \langle Mx, y \rangle.$$

Let g_m and g'_n be independent standard Gaussian vectors in $\mathbf{R}^m$ and $\mathbf{R}^n$.
Consider the Gaussian procceses indexed by $S^{m-1} \times S^{n-1}$ defined by
$X_{(x,y)} = \langle Mx, y \rangle$ and $Y_{(x,y)} = \langle g_m, x \rangle + \langle g'_n, y \rangle$.
Fact: $\|X_{(x,y)} - X_{(x',y')}\|_{L^2} \leqslant \|Y_{(x,y)} - Y_{(x',y')}\|_{L^2}$
Proof

Slepian's lemma implies that

$$\mathbf{E}\|M\|_{op} = \mathbf{E} \sup_{(x,y)} X_{(x,y)} \leqslant \mathbf{E} \sup_{(x,y)} Y_{(x,y)} \leqslant \sqrt{m} + \sqrt{n}.$$

This bound is very sharp! Simple check on Matlab gives

```
norm(randn(400,900))
ans = 49.5135
```

Next time: more on random matrices
How to use them for compressed sensing.