0.1 Sum-of-Squares Certifiably from the Poincaré inequality

We now turn our attention to bounding the value of (**??**). Ignoring finite-sample issues, our goal is to identify assumptions on p such that $M_{2k}(p) \stackrel{\text{def}}{=} \mathbb{E}_{X \sim p}[(X-\mu)^{\otimes 2k}]$ yields a small value for (**??**).

Before doing so, we will introduce some machinery for establishing bounds on (**??**). The main idea is that of a sum-of-squares proof:

Definition 0.1. A polynomial inequality $p(v) \leq q(v)$ has a *sum-of-squares proof* if $q(v) - p(v) \succeq_{\text{sos}} 0$. We will also denote this as $q(v) \succeq_{\text{sos}} p(v)$ or $p(v) \preceq_{\text{sos}} q(v)$.

The usefulness of this perspective is that the relation $\preceq_{\text{sos}}$ satisfies many of the same properties as $\leq$:

- If $p_1 \preceq_{\text{sos}} p_2$ and $p_2 \preceq_{\text{sos}} p_3$, then $p_1 \preceq_{\text{sos}} p_3$.
- If $p_1 \preceq_{\text{sos}} q_1$ and $p_2 \preceq_{\text{sos}} q_2$, then $p_1 + p_2 \preceq_{\text{sos}} q_1 + q_2$.
- If $p_1 \succeq_{\text{sos}} 0$ and $p_2 \succeq_{\text{sos}} 0$, then $p_1 p_2 \succeq_{\text{sos}} 0$.
- If $p_1 \preceq_{\text{sos}} p_2$, $q_1 \preceq_{\text{sos}} q_2$, and $p_2, q_1 \succeq_{\text{sos}} 0$, then $p_1 q_1 \preceq_{\text{sos}} p_2 q_2$.
- Moreover, many "standard" inequalities such as Cauchy-Schwarz and Hölder have sum-of-squares proofs.

Using these, we can often turn a normal proof that $p \leq q$ into a sum-of-squares proof that $p \preceq q$ as long as we give sum-of-squares proofs for a small number of key steps.

For concreteness, we will prove the last two claims properties above. We first prove that $p_1, p_2 \succeq_{\text{sos}} 0 \implies p_1 p_2 \succeq_{\text{sos}} 0$. Indeed we have

$$p_1(v)p_2(v) = (\sum_i p_{1i}(v)^2)(\sum_j p_{2j}(v)^2) = \sum_{ij} (p_1 i(v) p_{2j}(v))^2 \succeq_{\text{sos}} 0 \tag{1}$$

Next we prove that $p_1 \preceq_{\text{sos}} p_2, q_1 \preceq_{\text{sos}} q_2$, and $p_2, q_1 \succeq_{\text{sos}} 0$ implies $p_1 q_2 \preceq_{\text{sos}} p_2 q_2$. This is because

$$p_2 q_2 - p_1 q_1 = p_2(q_2 - q_1) + (p_2 - p_1) q_1 \succeq_{\text{sos}} 0, \tag{2}$$

where the second relation uses $p_2, q_2 - q_1 \succeq_{\text{sos}} 0$ and $p_2 - p_1, q_1 \succeq_{\text{sos}} 0$ together with the previous result.

In view of this, we can reframe bounding (**??**) as the following goal:

> **Goal:** Find a sum-of-squares proof that $\langle M_{2k}(p), v^{\otimes 2k} \rangle \preceq_{\text{sos}} \lambda \|v\|_2^{2k}$.

Certifiability for Gaussians. We now return to the assumptions needed on p that will enable us to provide the desired sum-of-squares proof. Let us start by observing that a sum-of-squares proof exists for any Gaussian distribution: If $p = \mathcal{N}(\mu, \Sigma)$, then

$$\begin{aligned}
\langle M_{2k}(\mathcal{N}(\mu, \Sigma)), v^{\otimes 2k} \rangle &= \langle M_{2k}(\mathcal{N}(0, I)), (\Sigma^{1/2} v)^{\otimes 2k} \rangle && (3)\\
&= \big(\prod_{i=1}^{k} (2i-1)\big) \langle \mathcal{I}, (\Sigma^{1/2} v)^{\otimes 2k} \rangle && (4)\\
&= \big(\prod_{i=1}^{k} (2i-1)\big) \|\Sigma^{1/2} v\|_2^{2k} && (5)\\
&\leq (2k)^k \|\Sigma\|^k \|v\|_2^{2k}, && (6)
\end{aligned}$$

so we may take $\lambda = (2k\|\Sigma\|)^k$. (Here $\mathcal{I}$ denotes the identity tensor that is 1 along the diagonal and zero elsewhere.) Therefore normal distributions have certifiably bounded moments, but the proof above heavily exploited the rotational symmetry of a normal distribution. We can provide similar proofs for other highly symmetric distributions (such as the uniform distribution on the hypercube), but these are unsatisfying as they only apply under very specific distributional assumptions. We would like more general properties that yield certifiably bounded moments.

Poincaré inequality. The property we will use is the *Poincaré inequality.* A distribution p on $\mathbb{R}^d$ is said to satisfy the Poincaré inequality with parameter σ if

$$\mathsf{Var}_{x\sim p}[f(x)] \leq \sigma^2 \mathbb{E}_{x\sim p}[\|\nabla f(x)\|_2^2] \tag{7}$$

for all differentiable functions $f : \mathbb{R}^d \to \mathbb{R}$. This is a "global to local property"–it says that for any function that for any function f that varies under p, that variation can be picked up in terms of local variation (the gradient). In particular, it says that p doesn't have any "holes" (regines with low probability density that lie between two regions of high probability density). Indeed, suppose that A and B were two disjoint convex regions with $p(A) = p(B) = \frac{1}{2}$. Then p cannot satisfy the Poincaré inequality with any constant, since there is a function that is 1 on A, 0 on B, and constant on both A and B.

Below are some additional examples and properties of Poincaré distributions:

- A one-dimensional Gaussian $\mathcal{N}(\mu, \sigma^2)$ is Poincaré with constant σ.
- If p, q are σ-Poincaré then their product $p \times q$ is σ-Poincaré. In particular a multivariate Gausssian $\mathcal{N}(\mu, \sigma^2 I)$ is σ-Poincaré.
- If $X \sim p$ is σ-Poincaré and A is a linear map, then AX is $(\sigma\|A\|)$-Poincaré. In particular, $aX_1 + aX_2$ is $(\sqrt{a^2+b^2}\sigma)$-Poincaré when X_1 and X_2 are both σ-Poincaré, and $\mathcal{N}(\mu, \Sigma)$ is $\|\Sigma\|^{1/2}$-Poincaré.
- More generally, if $X \sim p$ is σ-Poincaré and f is L-Lipschitz, then $f(X)$ is (σL)-Poincaré.

Together these imply that Poincaré distributions contain multivariate Gaussians, arbitrary Lipschitz functions of Gaussians, and independent sums of such distributions. The above properties (except the initial Gaussian property) are all straightforward computations. Let us next state two substantially deeper results:

- If p is σ-strongly log-concave (meaning that the log-probability density $\log p(x)$ satisfies $\nabla^2 \log p(x) \preceq -\frac{1}{\sigma^2} I$), then p is σ-Poincaré (**?**).
- Suppose that the support of $X \sim p$ has ℓ_2-radius at most R, and let $Z = \mathcal{N}(0, \tau^2 I)$ for $\tau \geq 2R$. Then $X + Z$ is $(\tau\sqrt{e})$-Poincaré (**?**).

Thus Poincaré encompasses all strongly log-concave densities, and effectively any product of bounded random variables (after adding Gaussian noise, which we can always do ourselves).

It is instructive to compare Poincaré to the sub-Gaussian property that we have so far relied on. Poincaré is neither strictly stronger or weaker than sub-Gaussian, but it is stronger than sub-exponential (we will see this below). In general, we should think of Poincaré as being substantially stronger than sub-exponential: it implies that not only is the distribution itself sub-exponential, but so is any Lipschitz function of the density.

As an example, consider the random variable $(X, Y) \in \mathbb{R}^d$ where $X \sim \mathcal{N}(0, I)$ and $Y = \epsilon X$ for a Rademacher random variable ϵ. Then (X, Y) is sub-Gaussian, but not Poincaré with good constant: if we take $f(X, Y) = \sum_i X_i Y_i$, then f is with high probability close to either $+d$ or $-d$, so $\mathsf{Var}[f(X, Y)] \approx d^2$. However, $\nabla f(X, Y) = (Y_1, \ldots, Y_d, X_1, \ldots, X_d)$ and so $\|\nabla f(X, Y)\|_2^2$ is close to $2d$ with high probability. Thus while the sub-Gaussian constant is $\mathcal{O}(1)$, the Poincaré constant in this case is $\Omega(\sqrt{d})$.

Consequences of Poincaré. So far we have seen conditions that imply Poincaré, but we would also like to derive consequences of this property. Below are some of the most useful ones:

- If $X \sim p$ is σ-Poincaré, then Lipschitz functions concentrate: $\mathbb{P}[|f(x) - \mathbb{E}[f(x)]| \geq t] \leq 6\exp(-t/(\sigma L))$ for any L-Lipschitz f.
- As a corollary, we have *volume expansion*: For any set A, let A_ϵ be the set of points within ℓ_2-distance ϵ of A. Then $p(A)p(A_\epsilon^c) \leq 36\exp(-\epsilon/\sigma)$.

This second property implies, for instance, that if $p(A) \geq \delta$, then almost all points will be within distance $\mathcal{O}(\sigma \log(1/\delta))$ of A.

To prove the second property, let $f(x) = \min(\inf_{y \in A} \|x - y\|_2 . \epsilon)$. Then f is Lipschitz, is 0 on A, and is ϵ on A_ϵ^c. Let μ be the mean of $f(X)$. Since f is sub-exponential we have $p(A) = p(f(X) = 0) \leq 6\exp(-\mu/\sigma)$, and $p(A_\epsilon^c) = p(f(X) = \epsilon) \leq 6\exp(-(\epsilon - \mu))/\sigma)$. Multiplying these together yields the claimed result.

The most important property for our purposes, however, will be the following:

Theorem 0.2. *Suppose that p is σ-Poincaré and let f be a differentiable function such that $\mathbb{E}[\nabla^j f(X)] = 0$ for $j = 1, \dots, k-1$. Then there is a universal constant C_k such that $\mathsf{Var}[f(X)] \leq C_k \sigma^{2k} \mathbb{E}[\|\nabla^k f(X)\|_F^2]$.*

Note that $k = 1$ is the original Poincaré property, so we can think of Theorem 0.2 as a generalization of Poincaré to higher derivatives. Note also that $\nabla^k f(X)$ is a tensor in $\mathbb{R}^{d^k}$; the notation $\|\nabla^k f(X)\|_F^2$ denotes the squared Frobenius norm of $\nabla^k f(X)$, i.e. the sum of the squares of its entries.

Theorem 0.2, while it may appear to be a simple generalization of the Poincaré property, is a deep result that was established in **?**, building on work of **?**. We will use Theorem 0.2 in the sequel to construct our sum-of-squares proofs.

Sum-of-squares proofs for Poincaré distributions. Here we will construct sum-of-squares proofs that $M_{2k}(v) \overset{\text{def}}{=} \mathbb{E}_p[\langle x - \mu, v \rangle^{2k}] \preceq_{\text{sos}} C'_k \sigma^{2k} \|v\|_2^{2k}$ whenever p is σ-Poincaré, for some universal constants C'_k. We will exhibit the proof for $k = 1, 2, 3$ (the proof extends to larger k and the key ideas appear already by $k = 3$). We introduce the notation

$$M_k = \mathbb{E}[(x - \mu)^{\otimes k}], \tag{8}$$

$$M_k(v) = \langle M_k, v^{\otimes k} \rangle = \mathbb{E}[\langle x - \mu, v \rangle^k]. \tag{9}$$

Proof for $k = 1$. We wish to show that $\mathbb{E}_p[\langle x - \mu, v \rangle^2] \preceq_{\text{sos}} \sigma^2 \|v\|_2^2$. To do this take $f_v(x) = \langle x, v \rangle$. Then the Poincaré inequality applied to f_v yields

$$\mathbb{E}_p[\langle x - \mu, v \rangle^2] = \mathsf{Var}[f_v(x)] \leq \sigma^2 \mathbb{E}[\|\nabla f_v(x)\|_2^2] = \sigma^2 \mathbb{E}[\|v\|_2^2] = \sigma^2 \|v\|_2^2. \tag{10}$$

Thus $M_2(v) \leq \sigma^2 \|v\|_2^2$ (this is just saying that Poincaré distributions have bounded covariance). This property has a sum-of-squares proof because it is equivalent to $\sigma^2 I - M_2 \succeq 0$, and we know that all positive semidefiniteness relations are sum-of-squares certifiable.

Proof for $k = 2$. Extending to $k = 2$, it makes sense to try $f_v(x) = \langle x - \mu, v \rangle^2$. Then we have $\nabla f_v(x) = 2\langle x - \mu, v \rangle v$ and hence $\mathbb{E}[\nabla f_v(x)] = 0$. We also have $\nabla^2 f_v(x) = 2v \otimes v$. Thus applying Theorem 0.2 we obtain

$$\mathsf{Var}[f_v(x)] \leq C_2 \sigma^4 \mathbb{E}[\|2v \otimes v\|_F^2] = 4C_2 \sigma^4 \|v\|_2^4. \tag{11}$$

We also have $\mathsf{Var}[f_v(x)] = \mathbb{E}[\langle x - \mu, v \rangle^4] - \mathbb{E}[\langle x - \mu, v \rangle^2]^2 = M_4(v) - M_2(v)^2$. Thus

$$M_4(v) = (M_4(v) - M_2(v)^2) + M_2(v)^2 \tag{12}$$

$$\leq 4C_2 \sigma^4 \|v\|_2^4 + \sigma^4 \|v\|_2^4 = (4C_2 + 1)\sigma^4 \|v\|_2^4. \tag{13}$$

This shows that the fourth moment is bounded, but how can we construct a sum-of-squares proof? We already have that $M_2(v)^2 \preceq_{\text{sos}} \sigma^4 \|v\|_2^4$ (by $0 \preceq_{\text{sos}} M_2(v) \preceq_{\text{sos}} \sigma^2 \|v\|_2^2$ and the product property). Therefore we focus on bounding $M_4(v) - M_2(v)^2 = \mathsf{Var}[f_v(x)]$.

For this we will apply Theorem 0.2 to a modified version of $f_v(x)$. For a matrix A, let $f_A(x) = (x - \mu)^\top A (x - \mu) = \langle A, (x - \mu)^{\otimes 2} \rangle$. Then $f_v(x) = f_A(x)$ for $A = vv^\top$. By the same calculations as above we have $\mathbb{E}[\nabla f_A(x)] = 0$ and $\nabla^2 f_A(x) = 2A$. Thus by Theorem 0.2 we have

$$\mathsf{Var}[f_A(x)] \leq C_2 \sigma^4 \mathbb{E}[\|2A\|_F^2] = 4C_2 \sigma^4 \|A\|_F^2. \tag{14}$$

On the other hand, we have $\mathsf{Var}[f_A(x)] = \langle M_4, A \otimes A \rangle - \langle M_2, A \rangle^2 = \langle M_4 - M_2 \otimes M_2, A \otimes A \rangle$. Thus (14) implies that

$$\langle M_4 - M_2 \otimes M_2, A \otimes A \rangle \leq 4C_2 \sigma^4 \|A\|_F^2. \tag{15}$$

Another way of putting this is that $M_4 - M_2 \otimes M_2$, when considered as a matrix in $\mathbb{R}^{d^2 \times d^2}$, is smaller than $4C_2 \sigma^4 I$ in the semidefinite ordering. Hence $4C_2 \sigma^4 I - (M_4 - M_2 \otimes M_2) \succeq 0$ and so $4C_2 \sigma^4 \|v\|_2^4 - \langle M_4 - M_2 \otimes M_2, v^{\otimes 4} \rangle \preceq_{\text{sos}} 0$, giving us our desired sum-of-squares proof. To recap, we have:

$$M_4(v) = (M_4(v) - M_2(v)^2) + M_2(v)^2 \tag{16}$$

$$\preceq_{\text{sos}} 4C_2 \sigma^4 \|v\|_2^4 + \sigma^4 \|v\|_2^4 = (4C_2 + 1)\sigma^4 \|v\|_2^4, \tag{17}$$

so we can take $C_2' = 4C_2 + 1$.

Proof for $k = 3$. Inspired by the $k = 1, 2$ cases, we try $f_v(x) = \langle x - \mu, v\rangle^3$. However, this choice runs into problems, because $\nabla f_v(x) = 3\langle x - \mu, v\rangle^2 v$ and so $\mathbb{E}[\nabla f_v(x)] = 3M_2(v)v \neq 0$. We instead should take

$$\begin{aligned} f_v(x) &= \langle x - \mu, v\rangle^3 - 3M_2(v)\langle x - \mu, v\rangle, \text{ which yields} && (18)\\ \mathbb{E}[\nabla f_v(x)] &= \mathbb{E}[3\langle x - \mu, v\rangle^2 v - 3M_2(v)v] = 0, && (19)\\ \mathbb{E}[\nabla^2 f_v(x)] &= \mathbb{E}[6\langle x - \mu, v\rangle(v \otimes v)] = 0, && (20)\\ \nabla^3 f_v(x) &= 6(v \otimes v \otimes v). && (21) \end{aligned}$$

Applying Theorem 0.2 to $f_v(x)$ yields

$$\mathsf{Var}[f_v(x)] \leq C_3\sigma^6 \|6(v \otimes v \otimes v)\|_F^2 = 36C_3\sigma^6\|v\|_2^6. \tag{22}$$

We can additionally compute

$$\begin{aligned} \mathsf{Var}[f_v(x)] &= \mathbb{E}[(\langle x - \mu, v\rangle^3 - 3M_2(v)\langle x - \mu, v\rangle)^2] - \mathbb{E}[\langle x - \mu, v\rangle^3 - 3M_2(v)\langle x - \mu, v\rangle]^2 && (23)\\ &= M_6(v) - 6M_2(v)M_4(v) + 9M_2(v)^3 - M_3(v)^2. && (24) \end{aligned}$$

Since our goal is to bound $M_6(v)$, we re-arrange to obtain

$$\begin{aligned} M_6(v) &= \mathsf{Var}[f_v(x)] + 6M_2(v)M_4(v) + M_3(v)^2 - 9M_2(v)^2 && (25)\\ &\leq 36C_3\sigma^6\|v\|_2^6 + 6(\sigma^2\|v\|_2^2)(C_2'\sigma^4\|v\|_2^4) + M_3(v)^2 + 0 && (26) \end{aligned}$$

We can also use Hölder's inequality to obtain $M_3(v)^2 \leq M_2(v)M_4(v)$, which yields an overall bound of $M_6(v) \leq (36C_3 + 12C_2')\sigma^6\|v\|_2^6$.

We now turn this into a sum-of-squares proof. We need to show the following four relations:

$$\begin{aligned} &(i)\ \mathsf{Var}[f_v(x)] \preceq_{\mathrm{sos}} 36C_3\sigma^6\|v\|_2^6, \quad (ii)\ M_2(v)M_4(v) \preceq_{\mathrm{sos}} (\sigma^2\|v\|_2^2)(C_2'\sigma^4\|v\|_2^4), && (27)\\ &(iii)\ M_3(v) \preceq_{\mathrm{sos}} M_2(v)M_4(v), \quad (iv)\ -9M_2(v)^2 \preceq_{\mathrm{sos}} 0. && (28) \end{aligned}$$

The relation (ii) again follows by the product property of $\preceq_{\mathrm{sos}}$, while $-9M_2(v)^2 \preceq_{\mathrm{sos}} 0$ is direct because $M_2(v)^2$ is already a square. We will show in an exercise that the Hölder inequality in (iii) has a sum-of-squares proof, and focus on (i).

The relation (i) holds for reasons analogous to the $k = 2$ case. For a symmetric tensor $A \in \mathbb{R}^{d^3}$, let $f_A(x) = \langle A, (x - \mu)^{\otimes 3} - 3M_2 \otimes (x - \mu)\rangle$. Then just as before we have $\mathbb{E}[\nabla f_A(x)] = 0$, $\mathbb{E}[\nabla^2 f_A(x)] = 0$, and so $\mathsf{Var}[f_A(x)] \leq 36C_3\sigma^6\|A\|_F^2$, which implies that[1]

$$M_6 - 6M_2 \otimes M_4 + 9M_2 \otimes M_2 \otimes M_2 - M_3 \otimes M_3 \preceq 36C_3\sigma^6 I, \tag{29}$$

and hence $\mathsf{Var}[f_v(x)] \preceq_{\mathrm{sos}} 36C_3\sigma^6\|v\|_2^6$ (again because semidefinite relations have sum-of-squares proofs).

In summary, we have $M_6(v) \preceq_{\mathrm{sos}} (36C_3 + 12C_2')\sigma^6\|v\|_2^6$, as desired.

Generalizing to higher k. For higher k the proof is essentially the same. What is needed is a function $f_v(x)$ whose first $k-1$ derivates all have zero mean. This always exists and is unique up to scaling by constants. For instance, when $k = 4$ we can take $f_v(x) = \langle x-\mu, v\rangle^4 - 6M_2(v)\langle x-\mu, v\rangle^2 - 4M_3(v)\langle x-\mu, v\rangle - M_4(v) + 6M_2(v)^2$. This appears somewhat clunky but is a special case of a combinatorial sum. For the general case, let $\mathcal{T}_k$ be the set of all integer tuples $(i_0, i_1, \dots)$ such that $i_0 \geq 0$, $i_s \geq 2$ for $s > 0$, and $i_0 + i_1 + \cdots = k$. Then the general form is

$$f_{v,k}(x) = \sum_{(i_0,\dots,i_r)\in\mathcal{T}_k} (-1)^r \binom{k}{i_0 \ \cdots \ i_r} \langle x - \mu, v\rangle^{i_0} M_{i_1}(v)M_{i_2}(v)\cdots M_{i_r}(v). \tag{30}$$

The motivation for this formula is that it is the solution to $\nabla f_{v,k}(x) = kf_{v,k-1}(x)v$. Using $f_{v,k}$, one can construct sum-of-squares proofs by applying Theorem 0.2 to the analogous $f_{A,k}$ function as before, and then use induction, the product rule, and Hölder's inequality as in the $k = 3$ case.

[1] Actually this is not quite true because we only bound $\mathsf{Var}[f_A(x)]$ for symmetric tensors A. What is true is that this holds if we symmetrize the left-hand-side of (29), which involves averaging over all ways of splitting M_2 and M_4 over the 3 copies of $\mathbb{R}^d$ in $\mathbb{R}^{d\times d\times d}$.