# A Proof of Slepian's Inequality

March 23, 2011

## 1 Introduction

Let $H$ be a real, separable, infinite dimensional Hilbert space. We will only need finite dimensional subspaces of $H$, but infinite dimensionality frees us of some notational complications, because every finite dimensional Hilbert space can be isometrically embedded in $H$. An $n$-tuple of points $\mathbf{x} = (x_1, ..., x_n) \in H^n$ will be called a *configuration*. The integer $n$ will be fixed in all the following.

With $\gamma$ we denote the unique centered $H$-valued Gaussian random variable whose covariance is the identity operator. If $(e_k)$ is an orthonormal basis of $H$ and $(\gamma_k)$ is an infinite sequence of independent standard normal variables we can write $\gamma = \sum \gamma_k e_k$.

We define a function $\Gamma : H^n \to \mathbb{R}$ on configurations by

$$\Gamma(\mathbf{x}) = \mathbb{E} \max_{i=1}^{n} \langle x_i, \gamma \rangle .$$

Slepian's lemma states that $\Gamma$ has the following monotonicity property:

**Theorem 1** *If two configurations* $\mathbf{x}$ *and* $\mathbf{y}$ *satisfy* $\|x_i - x_j\| \leq \|y_i - y_j\|$ *for all* $1 \leq i < j \leq m$*, then* $\Gamma(\mathbf{x}) \leq \Gamma(\mathbf{y})$*.*

Let $X_i$ and $X_i'$ be centered Gaussian random variables. Construction of two configurations $\mathbf{x}$ and $\mathbf{y} \in H^m$ such that $\langle x_i, x_j \rangle = \mathbb{E}[X_i X_j]$ and $\langle y_i, y_j \rangle = \mathbb{E}[Y_i Y_j]$ gives the following, perhaps more familiar version:
If $\mathbb{E}\left[(X_i - X_j)^2\right] \leq \mathbb{E}\left[(Y_i - Y_j)^2\right]$ for all $i < j$ then $\mathbb{E}\max_{i=1}^m X_i \leq \mathbb{E}\max_{i=1}^m Y_i$.

In the literature there are proofs ([3],[2],[4],[6]) where the hypotheses are augmented by the condition that the variances of the $X_i$ and the $Y_i$ be equal. From this weaker result the weaker conclusion $\mathbb{E}\max_{i=1}^m X_i \leq 2\mathbb{E}\max_{i=1}^m Y_i$ is then derived. Theorem 1 immediately implies these weaker versions, but, despite frequent citations, proofs are difficult to find, which is one of the motivations for these pages.

Another motivation is to state the Theorem in somewhat greater generality, combining the classical work of Lorenz [5] and Fernique. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called L-subadditive if

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n : \; f(\mathbf{x}) + f(\mathbf{y}) \geq f(\mathbf{x} \wedge \mathbf{y}) + f(\mathbf{x} \vee \mathbf{y}),$$

where $\wedge$ and $\vee$ denote the coordinatewise minimum and maximum respectively. Then we have

**Theorem 2** *Suppose that $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is a tempered distribution, L-subadditive and satistfies*

$$f\left(r_1 + t, ..., r_n + t\right) = f\left(r_1, ..., r_n\right) + h\left(t\right),$$

*where $h\left(-t\right) = -h\left(t\right)$. Then for any two configurations $\mathbf{x}$ and $\mathbf{y}$ satisfying $\left\Vert x_i - x_j\right\Vert \leq \left\Vert y_i - y_j\right\Vert$ for all $1 \leq i < j \leq m$ we have*

$$\mathbb{E}f\left(\left\langle x_1, \gamma\right\rangle, ..., \left\langle x_n, \gamma\right\rangle\right) \leq \mathbb{E}f\left(\left\langle y_1, \gamma\right\rangle, ..., \left\langle y_n, \gamma\right\rangle\right).$$

It turns out that the max-function is L-subadditive, so Theorem 1 follows from Theorem 2. Also the function $\text{diam}(x_1, ..., x_m) = \max_{i,j}\left(x_i - x_j\right)$ turns out to be L-subadditive, which leads to Ferniques result (Theorem 3.15 in [4]):

**Theorem 3** *If two configurations $\mathbf{x}$ and $\mathbf{y}$ satisfy $\left\Vert x_i - x_j\right\Vert \leq \left\Vert y_i - y_j\right\Vert$ for all $1 \leq i < j \leq m$, then*

$$\mathbb{E}\max_{i,j}\left\langle x_i - x_j, \gamma\right\rangle \leq \mathbb{E}\max_{i,j}\left\langle y_i - y_j, \gamma\right\rangle.$$

The proof of Theorem 2 is hinged on two independent results. The first one is that an analogous result holds if the function $f$ is twice differentiable and has non-positive mixed partial derivatives. I will establish this in the next section. The second result is the observation that the twice differentiable L-subadditive functions are characterized by exactly this property of having non-positive mixed partial derivatives. Combining the two results with a limiting argument leads to Theorem 2.

Slepian's inequality has important applications in the theory of empirical processes, such as the Sudakov minoration property (see e.g. [4]). More recently there have been applications in statistical learning theory, where it easily yields dimension-free uniform bounds for rather complicated function classes. Examples are the function classes of decision trees or neural networks [1],for $k$-means clustering, and various related techniques[7] and certain classes of operators involved in transfer learning [8]. The proof of Slepian's inequality is not particularly difficult, but it involves probability theory, linear algebra and analysis, and it is mathematically interesting.

## 2 Geometry

$H$ and $H^n$ have already been introduced. With $H_0^n$ we denote the set of all configurations $\mathbf{x} = (x_1, ..., x_n) \in H^n$ such that no $x_i$ is in the hyperplane generated by the other $x_j$, so that the $x_i$ form the vertices of a nondegenerate $n$-simplex.

A function $F : H^n \rightarrow \mathbb{R}$ is called *isometrically invariant* if it is invariant under translations and under unitary treansformations, that is

$$F\left(x_1 + y, ..., x_n + y\right) = F\left(x_1, ..., x_n\right) = F\left(Ux_1, ..., Ux_n\right)$$

for all $y \in H$ and unitary $U \in \mathcal{L}(H)$.

For a configuration $\mathbf{x} \in H^n$ let the vector $D(\mathbf{x}) \in \mathbb{R}^{n(n-1)/2}$ be defined by

$$D(\mathbf{x})_{ij} = \|x_i - x_j\|^2 \text{ for } 1 \leq i < j \leq n.$$

In the following we denote $\Delta = D(H^n)$ and $\Delta_0 = D(H_0^n)$. If $a \in \Delta$ and $a = D(\mathbf{x})$, then the configuration $\mathbf{x}$ is called a realization of $a$. A function $g : \Delta \to \mathbb{R}$ is called nondecreasing if $g(a) \leq g(b)$ whenever $a_{ij} \leq b_{ij}$ for all $1 \leq i < j \leq n$.

**Lemma 4** *$\Delta$ and $\Delta_0$ are convex, $\Delta_0$ is open and $\Delta$ is the closure of $\Delta_0$.*

*Suppose that $F : H^n \to \mathbb{R}$ is isometrically invariant. Then there exists $g : \Delta \to \mathbb{R}$ such that*

$$F = g \circ D.$$

*Furthermore, if $F$ is continuous then $g$ is continuous, and if $F$ is differentiable then $g$ is differentiable in $\Delta_0$.*

**Proof.** Let $a \in \mathbb{R}^{n(n-1)/2}$ and define a symmetric $(n-1) \times (n-1)$-matrix $G(a)$ by $G(a)_{ij} = (\bar{a}_{ni} + \bar{a}_{nj} - \bar{a}_{ij})/2$, where $\bar{a}$ denotes the extension of $a$ to a symmetric matrix vanishing on the diagonal. If $a \in \Delta$ and $\mathbf{x}$ is any realization of $a$ denote with $\mathbf{x}_0$ the realization $\mathbf{x}_0 = (x_1 - x_n, ..., x_{n-1} - x_n, 0)$. Then for $i, j \in \{1, ..., n-1\}$ we have

$$G(a)_{ij} = \langle x_i - x_n, x_j - x_n \rangle,$$

so $G(a)$ is the gramian of the first $n-1$ vectors is $\mathbf{x}_0$. It follows that $G(a)$ is positive semidefinite if $a \in \Delta$ and positive definite if $a \in \Delta_0$. Since $a \mapsto G(a)$ is continuous $\Delta_0$ is open. Let $(e_k)$ be a standard orthonormal basis of $H$ and $G^{1/2}(a)$ the positive semidefinite square-root of $G(a)$. We define $\mathbf{x}_s(a)$ by

$$\mathbf{x}_s(a) = \left( \sum_{i=1}^{n-1} G^{1/2}(a)_{1i} e_i, ..., \sum_{i=1}^{n-1} G^{1/2}(a)_{(n-1)i} e_i, 0 \right).$$

It is easily checked that $\mathbf{x}_s(a)$ is a realization of $a$, which we call the standard realization. Observe that $a \mapsto G(a)$ is differentiable and that $G(a) \mapsto G^{1/2}(a)$ is continuous and also differentiable if $G(a)$ is positive definite. It follows that $a \mapsto \mathbf{x}_s(a)$ is continuous and for $a \in \Delta_0$ also differentiable. If $\lambda \in [0, 1]$ and $a_1, a_2 \in \Delta$, then $G((1-\lambda) a_1 + \lambda a_2) = (1-\lambda) G(a_1) + \lambda G(a_2)$ is positive definite, so by the above construction $(1-\lambda) a_1 + \lambda a_2 \in \Delta$, so that $\Delta$ is convex. It follows similarly that $\Delta_0$ is convex. Define

$$g(a) = F(\mathbf{x}_s(a)), \text{ for } a \in \Delta,$$

which gives the required continuity and differentiability properties of $g$ and also shows that $\Delta$ is the closure of $\Delta_0$. It remains to verify that $F = g \circ D$, that is $F(\mathbf{x}) = F(\mathbf{x}_s(a))$ for any realization $\mathbf{x}$ of $a$. Clearly translation by $x_n$ is an isometry, so $F(\mathbf{x}) = F(\mathbf{x}_0)$. Let $(f_k)$ be any orthonormal basis of $H$ and $A$ an

$(n-1)\times(n-1)$-matrix such that $x_i - x_n = \sum_{j=1}^{n-1} A_{ij} f_j$ . Then $A = UG^{1/2}(a)$ for orthogonal $U$, by polar decomposition, so $\sum_j \beta_j f_j \mapsto \sum_{jk} U_{jk}\beta_j e_k$ is an isometry taking $\mathbf{x}_0$ to $\mathbf{x}_s(a)$, whence $F(\mathbf{x}) = F(\mathbf{x}_0) = F(\mathbf{x}_s(a))$. ■

For a function $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2$, we denote with $\partial_k f$ and $\partial_{ik} f$ the second partial derivative w.r.t. the $k$-th and $i$-th variables.

**Theorem 5** *If $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^2$, $(\partial_{ik} f) \leq 0$ for $i \neq k$ and*

$$f(r_1 + t, ..., r_n + t) = f(r_1, ..., r_n) + g(t),$$

*where $g(-t) = -g(t)$, then the configuration functional*

$$\Phi_f(\mathbf{x}) = \mathbb{E} f(\langle x_1, \gamma\rangle, ..., \langle x_n, \gamma\rangle) \text{ for } \mathbf{x} \in H^n$$

*is isometrically invariant and satisfies*

$$\Phi_f(\mathbf{x}) \leq \Phi_f(\mathbf{y}) \text{ if } \|x_i - x_j\|^2 \leq \|y_i - y_j\|^2 \text{ for all } 1 \leq i, j \leq n.$$

*nondecreasing.*

**Proof.** Orthogonal invariance of $\Phi$ is an obvious consequence of the orthogonal invariance of the gaussian expectation, and translation invariance follows because $g$ is odd and the expectation is invariant under the transformation $\gamma \mapsto -\gamma$. This gives isometric invariance.

Differentiability of $f$ implies that $\Phi_f$ is differentiable. By isometric invariance and the previous lemma there must be a continuous function $g$ on $\Delta$, differentiable on $\Delta_0$, such that $g \circ D = \Phi_f$. We need to show that $g$ is nondecreasing. Since $g$ is continuous and $\Delta$ is the closure of $\Delta_0$ it is enough to verify this on $\Delta_0$ and, since $g$ is differentiable on the convex set $\Delta_0$ it suffices to show that all partial derivatives of $g$ are non-negative.

Fix $a \in \Delta_0$ and indices $1 \leq k < l \leq n$. Let $a = D(\mathbf{x}) = D(x_1, ..., x_n)$, where we can assume $x_k = 0$ by translation invariance. Since $a \in \Delta_0$, the $x_i$ are linearly independent for $i \neq l$ and we can write $x_l = z - y$, where $z \in Span\{x_i : i \neq l\}$, and $y \perp Span\{x_i : i \neq l\}$ and $\|y\| > 0$. Define a curve $\gamma : \mathbb{R} \to H^n$ by

$$\gamma(t) = (x_1(t), ..., x_n(t)) = (x_1, ..., x_{k-1}, ty, x_{k+1}, ..., x_n).$$

so only the $k$-th component is made time-dependent and set to $ty$. Defining $\eta : \mathbb{R} \to \Delta$ by $\eta(t) = D(\gamma(t))$ we find for $1 \leq i < j \leq n$ that

$$\dot{\eta}_{ij}(0) = \left(\frac{d}{dt}\right)_{t=0} \|x_i(t) - x_j(t)\|^2 = \begin{cases} 2\|y\|^2 & \text{if } i = k \text{ and } j = l \\ 0 & \text{otherwise} \end{cases}.$$

We therefore have

$$2\|y\|^2 \frac{\partial g}{\partial e^{(kl)}}(a) = \sum_{i<j} \frac{\partial g}{\partial e^{(ij)}} \dot{\eta}_{ij}(0) = \frac{d}{dt} g(\eta(t))(0) = \frac{d}{dt} \Phi_f(\gamma(t))(0), \quad (1)$$

where $e^{(kl)}$ denotes the basis vector of $\mathbb{R}^{n(n-1)/2}$ given by $e_{ij}^{(kl)} = 1$ if $i = k$ and $j = l$ and zero else.

We now write any $\gamma \in H$ as $\gamma = \gamma_0 + \gamma^{\perp}$ where $\gamma_0 = \langle y, \gamma \rangle y / \|y\|^2$ and $\gamma^{\perp}$ is orthogonal to $y$. Using the shorthand $(\mathbf{x}, \gamma) = (\langle x_1, \gamma \rangle, ..., 0, ... \langle x_m, \gamma \rangle)$ we define a function $h_{\gamma}(t) = (\partial_k f)\left(\mathbf{x}, \gamma^{\perp} + t\gamma_0\right)$. Observe that $\langle x_i, y \rangle = 0$ for $i \neq l$ and $\langle x_l, y \rangle = -\|y\|^2 \leq 0$. Then

$$\begin{aligned} h'_{\gamma}(t) \langle y, \gamma_0 \rangle &= \sum_{i \neq k} (\partial_{ik} f)\left(\mathbf{x}, \gamma^{\perp} + t\gamma_0\right) \langle x_i, \gamma_0 \rangle \langle y, \gamma_0 \rangle \\ &= (\partial_{lk} f)\left(\mathbf{x}, \gamma^{\perp} + t\gamma_0\right) \langle x_l, \gamma_0 \rangle \langle y, \gamma_0 \rangle \\ &= -(\partial_{lk} f)\left(\mathbf{x}, \gamma^{\perp} + t\gamma_0\right) \langle y, \gamma \rangle^2 \geq 0, \end{aligned}$$

where we used the hypothesis on the mixed partials in the last step.

For $\gamma \in H$ denote $\hat{\gamma} = -\gamma_0 + \gamma^{\perp}$, corresponding to a reflection on the hyperplane orthogonal to $y$. The expectation $\mathbb{E}_{\gamma}$ is invariant under the reflection $\gamma \to \hat{\gamma}$. So by (1) and $h'_{\gamma}(t) \langle y, \gamma_0 \rangle \geq 0$

$$\begin{aligned} 2\|y\|^2 \frac{\partial g}{\partial e^{(kl)}}(a) &= \frac{d}{dt} \Phi_f(x_1, .., ty, ..., x_m)(0) \\ &= \mathbb{E} \langle y, \gamma \rangle (\partial_k f)(\mathbf{x}, \gamma) \\ &= \frac{1}{2} \left(\mathbb{E} \langle y, \gamma \rangle (\partial_k f)(\mathbf{x}, \gamma) + \mathbb{E} \langle y, \hat{\gamma} \rangle (\partial_k f)(\mathbf{x}, \hat{\gamma})\right) \\ &= \frac{1}{2} \left(\mathbb{E} \langle y, \gamma_0 \rangle (\partial_k f)\left(\mathbf{x}, \gamma^{\perp} + \gamma_0\right) - \mathbb{E} \langle y, \gamma_0 \rangle (\partial_k f)\left(\mathbf{x}, \gamma^{\perp} - \gamma_0\right)\right) \\ &= \frac{1}{2} \mathbb{E} \langle y, \gamma_0 \rangle (h_{\gamma}(1) - h_{\gamma}(-1)) \\ &= \frac{1}{2} \mathbb{E} \int_{-1}^{1} h'_{\gamma}(t) \langle y, \gamma_0 \rangle \, dt \geq 0. \end{aligned}$$

So $\partial g / \partial e^{(kl)}$ is nonnegative in $\Delta_0$ which implies the required monotonicity of $g$ on the convex set $\Delta_0$. Continuity of $g$ extends this property to all of $\Delta$. ■

Clearly the monotonicity property extends to more general functions which are the pointwise limits of $C^2$-functions with non-positive mixed partial derivatives. In the next section we characterize this class of functions.

# 3 L-subadditive functions

This concept, together with the symmetrically defined L-superadditive functions) was introduced by Lorentz [5] in 1953. For two real numbers $s$ and $t$ we use $s \wedge t$ and $s \vee t$ to denote respectively their minimum and maximum.

**Definition 6** *For $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ the minimum $\mathbf{x} \wedge \mathbf{y} \in \mathbb{R}^m$ and the maximum $\mathbf{x} \vee \mathbf{y} \in \mathbb{R}^m$ are defined by*

$$\begin{aligned} (\mathbf{x} \wedge \mathbf{y}) &= (x_1 \wedge y_1, ..., x_m \wedge y_m) \\ (\mathbf{x} \vee \mathbf{y}) &= (x_1 \vee y_1, ..., x_m \vee y_m). \end{aligned}$$

Note the translation invariance properties

$$\begin{aligned}(\mathbf{x}+\mathbf{z})\wedge(\mathbf{y}+\mathbf{z}) &= (\mathbf{x}\wedge\mathbf{y})+\mathbf{z}\\ (\mathbf{x}+\mathbf{z})\vee(\mathbf{y}+\mathbf{z}) &= (\mathbf{x}\vee\mathbf{y})+\mathbf{z}.\end{aligned}$$

**Definition 7** *A function $f:\mathbb{R}^m\to\mathbb{R}$ is called L-subadditive if*

$$\forall\mathbf{x},\mathbf{y}\in\mathbb{R}^m:\ f(\mathbf{x})+f(\mathbf{y})\geq f(\mathbf{x}\wedge\mathbf{y})+f(\mathbf{x}\vee\mathbf{y}).$$

**Lemma 8** *Every locally integrable L-subadditive function $f:\mathbb{R}^m\to\mathbb{R}$ is the pointwise limit of infinitely differentiable L-subadditive functions $f_n$. If $f$ is continuous the convergence is uniform on compact subsets of $\mathbb{R}^m$. If $f$ has the property that*

$$f(r_1+t,...,r_n+t)=f(r_1,...,r_n)+g(t),$$

*where $g(-t)=-g(t)$, then $f_n$ can be chosen to have the same property with some appropriate $g_n$.*

**Proof.** Choose a suitable sequence of nonnegative test functions $h_n$ such that such that $f_n=f*h_n\to f$ pointwise (or uniformly on compact subsets if $f$ is continuous). Since $f*h_n$ is $C^\infty$ it suffices to show that $f*h$ is L-subadditive for every positive measurable $g\in L_1$. But

$$\begin{aligned}f*h(\mathbf{x})+f*h(\mathbf{y}) &= \int_{\mathbb{R}^m}(f(\mathbf{x}-z)+f(\mathbf{y}-z))\,h(z)\,dz\\ &\geq \int_{\mathbb{R}^m}(f(\mathbf{x}\vee\mathbf{y}-z)+f(\mathbf{x}\wedge\mathbf{y}-z))\,h(z)\,dz\\ &= f*gh\mathbf{x}\wedge\mathbf{y}+f*h(\mathbf{x}\vee\mathbf{y}).\end{aligned}$$

This proves the first part. For the second it is easily verified that

$$f_n(r_1+t,...,r_n+t)=f_n(r_1,...,r_n)+g(t)\left(\int_{\mathbb{R}^m}h_n(z)\,dz\right).$$

■

Below I denote $\partial_i$ for $\partial/\partial r_i$ and $\partial_{ij}$ for $\partial^2/\partial r_i\partial r_j$, where the partial derivatives are understood in the sense of distributions. The L-subadditive functions are completely characterized by the sign of mixed second partial derivatives:

**Theorem 9** *If $f:\mathbb{R}^m\to\mathbb{R}$ is locally integrable then $f$ is L-subadditive if and only if $\partial_{kl}f\leq 0$, in the sense of distributions, for all $k\neq l$.*

**Proof.** Suppose first that $f\in C^2$. Suppose that $f$ is L-subadditive and let $\mathbf{x}\in\mathbb{R}^m$, $k,l\in\{1,...,m\}$. Since $f\in C^2$ we have (suppressing the dependence of

$f$ on other coordinates than $x_k$ and $r_l$)

$$
\begin{aligned}
&\left(\partial_{kl} f\right)(\mathbf{x}) \\
&= \lim_{0<s,t\to 0} (st)^{-1}\left(f\left(x_k+s, x_l+t\right) - f\left(x_k, x_l+t\right) - \left(f\left(x_k+s, x_l\right) - f\left(x_k, x_l\right)\right)\right) \\
&= \lim_{0<s,t\to 0} (st)^{-1}\left(f\left(x_k+s\vee x_k, x_l\vee x_l+t\right) + f\left(x_k+s\wedge x_k, x_l\wedge x_l+t\right)\right. \\
&\left. - \left(f\left(x_k, x_l+t\right) + f\left(x_k+s, x_l\right)\right)\right) \\
&\leq 0,
\end{aligned}
$$

which proves the only-if part.

Suppose now that $\partial_{kl} f \leq 0$ for all $k \neq l$ and let $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^m$. Define curves $\xi$ and $\eta : [0,1] \to \mathbb{R}^m$ by

$$
\begin{aligned}
\boldsymbol{\xi}(t) &= (1-t)\,\mathbf{x} + t\left(\mathbf{x}\wedge\mathbf{y}\right) \\
\boldsymbol{\eta}(t) &= (1-t)\left(\mathbf{x}\vee\mathbf{y}\right) + t\mathbf{y}.
\end{aligned}
$$

Let $I \subseteq \{1,...,m\}$ be the set of indices such that $x_i > y_i$. Observe that for $i \in I$ we have

$$\xi_i(t) = \eta_i(t) = (1-t)\,x_i + ty_i,$$

while for $i \notin I$ we have $\xi_i(t) = x_i$ and $\eta_i(t) = y_i$ for all $t$. Now define a real function $h$ by $h(t) = f(\boldsymbol{\xi}(t)) - f(\boldsymbol{\eta}(t))$. To show that $f$ is L-subadditive it suffices to show that $h(1) \leq h(0)$, which will follow if we can show that $h'(t) \leq 0$. Fix any $t$. We have

$$h'(t) = \sum_{i\in I}\left[\partial_i f(\boldsymbol{\eta}(t)) - \partial_i f(\boldsymbol{\xi}(t))\right](x_i - y_i).$$

To show that this is negative we define a curve $\boldsymbol{\chi} : [0,1] \to \mathbb{R}^m$ by $\boldsymbol{\chi}(s) = (1-s)\,\boldsymbol{\xi}(t) + s\boldsymbol{\eta}(t)$. Observe that

$$\chi_i(s) = \begin{cases} (1-t)\,x_i + ty_i & \text{if} \quad i \in I \\ (1-s)\,x_i + sy_i & \text{if} \quad i \notin I \end{cases}.$$

We therefore obtain

$$
\begin{aligned}
h'(t) &= \sum_{i\in I}\left[\partial_i f(\boldsymbol{\chi}(1)) - \partial_i f(\boldsymbol{\chi}(0))\right](x_i - y_i) \\
&= \sum_{i\in I}\int_0^1 \frac{d}{ds}\partial_i f(\boldsymbol{\chi}(s))\,ds\,(x_i - y_i) \\
&= \int_0^1 \sum_{i\in I}\sum_{j\notin I}\partial_{ij} f(\boldsymbol{\chi}(s))\,(y_j - x_j)\,(x_i - y_i)\,ds \\
&\leq 0.
\end{aligned}
$$

The inequality holds because $(y_j - x_j)(x_i - y_i) \geq 0$ for $i \in I$ and $j \notin I$.

We have shown that the announced result holds for $f \in C^2$. Now let $k \neq l$ and let $f$ be locally integrable and $h_n$ and $f_n$ as in Lemma 8. Since

$$\partial_{kl} f_n = \partial_{kl} \left( f * h_n \right) = \left( \partial_{kl} f \right) * h_n$$

we have $\left( \partial_{kl} f \right) \leq 0 \implies \partial_{kl} f_n \leq 0 \implies f_n$ is L-subadditive. Since L-subadditivity is preserved under pointwise limits this implies that $f$ is L-subadditive. Conversely, by the previous Lemma, if $f$ is L-subadditive then so is $f_n$, whence $\partial_{kl} f_n \leq 0 \implies \partial_{kl} f \leq 0$. ■

# 4 Putting it together

**Proof of Theorem 2.** By Lemma 8 we can find a sequence $f_n$ of L-subadditive $C^\infty$-functions such that $f_n \to f$ and

$$f_n \left( r_1 + t, ..., r_n + t \right) = f_n \left( r_1, ..., r_n \right) + h_n \left( t \right),$$

where $h_n \left( -t \right) = -h_n \left( t \right)$. By Theorem 9 we have $\partial_{kl} f_n \leq 0$ for all $k \neq l$, so that $f_n$ satisfies the hypotheses of Theorem 5. $\Phi_{f_n}$ has therefore the required monotonicity property, which is preserved under the limit $f_n \to f$, because $f$ and $f_n$ are tempered distributions integrated against Gaussians. ■

**Proof of Theorem 1.** The maximum function $\left( x_1, ..., x_m \right) \mapsto \max \left\{ x_1, ..., x_m \right\}$ is L-subadditive: Take $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$, wlog assume $\max \left( \mathbf{x} \right) \geq \max \left( \mathbf{y} \right)$. Then $\max \left( \mathbf{x} \vee \mathbf{y} \right) \leq \max \left( \mathbf{x} \right)$ and $\max \left( \mathbf{x} \wedge \mathbf{y} \right) \leq \max \left( \mathbf{y} \right)$ and adding these inequalities shows L-subadditivity. We also have

$$\max \left\{ x_1 + t, ..., x_n + t \right\} = \max \left\{ x_1, ..., x_n \right\} + t,$$

so $f = \max$ satisfies the hypotheses of Theorem 2 with $g \left( t \right) = t$. ■

**Proof of Theorem 3.** Consider the function

$$\operatorname{diam} \left( x_1, ..., x_m \right) = \max_{i,j} \left( x_i - x_j \right).$$

Take $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ and consider the quantity

$$A_{ijkl} = x_i \vee y_i - x_j \vee y_j + x_i \wedge y_i - x_j \wedge y_j.$$

Assume first $x_i \geq y_i$. Then if $y_l < x_l \implies A_{ijkl} \leq x_i - x_j + y_k - y_l \leq$diam$(\mathbf{x})$ +diam$(\mathbf{y})$, and $y_l \geq x_l \implies A_{ijkl} \leq x_i - x_l + y_k - y_j \leq$diam$(\mathbf{x})$ +diam$(\mathbf{y})$. Analogous inequalities hold if $x_i < y_i$, so that in all cases $A_{ijkl} \leq$diam$(\mathbf{x})$ +diam$(\mathbf{y})$. Maximizing over all pairs $(i,j)$ and $(k,l)$ shows that diam is L-subadditive. It is also invariant under translations.

Now let $\psi$ be as in the theorem. We find

$$\partial_{kl} \left( \psi \left( \operatorname{diam} \left( \mathbf{x} \right) \right) \right) = \psi'' \left( \operatorname{diam} \left( \mathbf{x} \right) \right) \left( \partial_k \operatorname{diam} \right) \left( \mathbf{x} \right) \left( \partial_l \operatorname{diam} \right) \left( \mathbf{x} \right) + \psi' \left( \operatorname{diam} \left( \mathbf{x} \right) \right) \left( \partial_{kl} \operatorname{diam} \right) \left( \mathbf{x} \right).$$

The second term on the right is non-positive because $\psi$ is nondecreasing and by L-subadditivity of diam combined with by Theorem 9. By convexity of $\psi$ we get $\psi''(\mathrm{diam}(\mathbf{x})) \geq 0$. But the product $(\partial_k \mathrm{diam})(\mathbf{x})(\partial_l \mathrm{diam})(\mathbf{x})$ is nonzero only if either $\mathrm{diam}(\mathbf{x}) = x_k - x_l$ or $\mathrm{diam}(\mathbf{x}) = x_l - x_k$. In both cases the prosuct is negative, so $\partial_{kl}(\psi \circ \mathrm{diam}) \leq 0$. So $\psi \circ \mathrm{diam}$ is L-subadditive and satisfies the hypotheses of 2 with $g(t) = 0$. ■

# References

[1] P. L. Bartlett and S. Mendelson. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. *Journal of Machine Learning Research*, 3: 463–482, 2002.

[2] F. W. Huffer. Slepian's Inequality via the central limit theorem, *The Canadian Journal of Statistics*, Vol. 14, No. 4, pp. 367-370, 1986

[3] K. Joag-Dev, M. D. Perlman, L. D. Pitt. Association of normal random variables and Slepian's inequality. *The Annals of Probability*, Vol. 11, No. 2, pp. 451-455

[4] M. Ledoux, M. Talagrand. *Probability in Banach Spaces*, Springer, 1991.

[5] G. G. Lorentz. An inequality for rearrangements. Amer. Math. Monthly, 60, 176-179, 1953

[6] P. Massart. Concentration inequalities and model selection. Available online

[7] A. Maurer, M. Pontil, Uniform error bounds for K-dimensional coding schemes in Hilbert spaces, *ALT* 2008.

[8] A. Maurer, Transfer bounds for linear feature learning, *Machine Learning* 75,3, 327 - 350, 2009

[9] D. Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Tech. J.* 41, 463-501, 1962