# Every decision tree has an influential variable

Ryan O'Donnell
Microsoft Research*
odonnell@microsoft.com

Michael Saks
Rutgers
saks@math.rutgers.edu

Oded Schramm
Microsoft Research
schramm@microsoft.com

Rocco A. Servedio†
Columbia University
rocco@cs.columbia.edu

**Abstract**

In this paper we prove a new inequality relating the decision tree complexity and the influences of boolean functions. In particular, we show that any balanced boolean function with a decision tree of depth $d$ has a variable with influence at least $\frac{1}{d}$. The only previous nontrivial lower bound known was $\Omega(d2^{-d})$. Our inequality has many generalizations, allowing us to prove influence lower bounds for randomized decision trees, decision trees on arbitrary product probability spaces, and decision trees with non-boolean outputs. As an application of our results we give a very easy proof that the randomized query complexity of nontrivial monotone graph properties is at least $\Omega(v^{4/3}/p^{1/3})$, where $v$ is the number of vertices and $p \leq \frac{1}{2}$ is the critical threshold probability. This supersedes the milestone $\Omega(v^{4/3})$ bound of Hajnal [14] and is sometimes superior to the best known lower bounds of Chakrabarti-Khot [9] and Friedgut-Kahn-Wigderson [11].

*Some of this research was performed while this author was at the Institute for Advanced Study
†Supported in part by NSF CAREER award CCF-0347282 and a Sloan Foundation Fellowship.

# 1 Introduction

**1.1 Motivation.** This paper lies at the intersection of two topics within the theory of boolean functions.

The first topic is *decision tree complexity*. A *deterministic decision tree* (DDT) for a boolean function $f : \{-1,1\}^n \to \{-1,1\}$ is a deterministic adaptive strategy for reading variables so as to determine the value of $f$ (a formal definition appears in Section 3.1). The cost of a DDT on a given input is simply the number of input variables that it reads, and the DDT complexity of a function $f$, $D(f)$, is the minimum over all DDT's for $f$ of the maximum cost of any input. A *randomized decision tree* (RDT) for $f$ is a probability distribution over DDTs for $f$; such trees are sometimes known as *zero-error* randomized decision trees. The RDT complexity of $f$, $R(f)$, is the minimum over all RDT's for $f$ of the maximum expected cost of any input. Decision tree complexity has been studied in theoretical science for over 30 years and there is now a significant body of research on the subject (for a survey, see e.g. [8]).

The second topic is *variable influences*, introduced to theoretical computer science by Ben-Or and Linial in 1985 [2]. Any $n$-variate boolean function $f$ has an associated *influence vector* $(\mathbf{Inf}_1(f), \ldots, \mathbf{Inf}_n(f))$ where $\mathbf{Inf}_i(f)$ measures the extent to which the value of $f$ depends on variable $i$ (a precise definition appears in Section 1.2). A number of papers have dealt with properties of this vector and its relation to other properties of boolean functions; perhaps the best known work along these lines is that of Kahn, Kalai and Linial [15] ("KKL") concerning the maximum influence $\mathbf{Inf}_{\max}(f) = \max\{\mathbf{Inf}_i(f) : i \in [n]\}$. Their result implies, for example, that $\mathbf{Inf}_{\max}(f) = \Omega(\frac{\log n}{n})$ for any near-balanced boolean function $f$ (where we say that $f$ is near-balanced if both $|f^{-1}(1)|/2^n$ and $|f^{-1}(-1)|/2^n$ are $\Omega(1)$).

The question that originally motivated this paper was: what is the best lower bound on $\mathbf{Inf}_{\max}(f)$ that holds for all near-balanced boolean functions $f$ satisfying $D(f) \leq d$? It is easy to see that such a function $f$ depends on at most $2^d$ of its variables and therefore the KKL result implies $\mathbf{Inf}_{\max}(f) \geq \Omega(\frac{d}{2^d})$; prior to this work, this was the best lower bound known. Our main inequality for boolean functions, Theorem 1.1, implies a (tight) lower bound of $\mathbf{Inf}_{\max}(f) \geq \Omega(\frac{1}{d})$ for any near-balanced function $f$ satisfying $D(f) \leq d$.

In fact, Theorem 1.1 provides a lower bound on a weighted average of the influence vector, where $\mathbf{Inf}_i(f)$ is weighted by the probability that a DDT for $f$ queries $x_i$ when $x$ is a randomly chosen *input*. This lets us extend our lower bound on $\mathbf{Inf}_{\max}(f)$ to functions with $R(f) \leq d$ and even to functions with $\Delta(f) \leq d$, where $\Delta(f)$ denotes the expected number of queries made by the best DDT for $f$ on a random input (again, see Section 1.2 for precise definitions).

**1.2 The main theorem for boolean functions.** Our main theorem holds in a very general setting, that of functions from product probability spaces into metric spaces. However the case of greatest interest to us is much simpler. Let $\{-1,1\}^n_{(p)}$ denote the discrete cube endowed with the $p$-biased measure, $0 < p < 1$; when we write simply $\{-1,1\}^n$ the uniform measure case $p = \frac{1}{2}$ is implied. Our main interest is in boolean functions $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$, and in this section we will describe our main theorem in this case.

First we recall a few definitions. We have

$$\mathbf{Var}[f] = \mathbf{E}[f^2] - \mathbf{E}[f]^2 = 4\mathbf{Pr}[f = 1]\mathbf{Pr}[f = -1].$$

This measures the "balance" of $f$; if $f$ is equally likely to be true or false then $\mathbf{Var}[f] = 1$. We also make the following definition for the *influence of the $i$th coordinate* on $f$:

$$\mathbf{Inf}_i(f) = 2 \mathop{\mathbf{Pr}}_{x,x^{(i)}}[f(x) \neq f(x^{(i)})],$$

where $x$ is drawn from $\{-1,1\}^n_{(p)}$ and $x^{(i)}$ is formed by *rerandomizing* the $i$th coordinate of $f$. Note that our definition agrees with the one introduced in [2] in the uniform measure case $p = \frac{1}{2}$, which was $\mathbf{Inf}_i[f] = \mathbf{Pr}[f(x) \neq f(x \oplus i)]$. (Our definition differs from the $p$-biased notion of influences used in, e.g., [12] by a factor of $4p(1-p)$; we prefer rerandomizing the $i$th coordinate to flipping it since this makes sense in more general product probability spaces which we will consider later.) We call $\mathbf{Inf}(f) := \sum_{i=1}^n \mathbf{Inf}_i(f)$ the *total influence* of $f$.

Finally, since the notion of influences involves randomizing over the input domain, it makes sense to introduce a notion of randomizing over inputs for decision trees. Let $T$ be a DDT computing a function $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$. We write

$$\delta_i^{(p)}(T) = \mathop{\mathbf{Pr}}_{x \in \{-1,1\}^n_{(p)}}[T \text{ queries } x_i], \qquad \text{and}$$

$$\Delta^{(p)}(T) = \sum_{i=1}^n \delta_i^{(p)}(T) = \mathop{\mathbf{E}}_{x \in \{-1,1\}^n_{(p)}}[\text{number of coordinates of } x \text{ queried by } T].$$

We also let $\Delta^{(p)}(f)$ denote the minimum of $\Delta^{(p)}(T)$ over all DDTs $T$ computing $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$. It is easy to see this is equivalent to minimizing over all RDTs computing $f$; hence $\Delta^{(p)}(f) \leq R(f)$ for all $p$. Also note that $\Delta^{(p)}(f)$ can be upper-bounded in terms of the *size* (number of leaves) of the smallest DDT for $f$: [20] shows $\Delta^{(p)}(f) \leq \log_2(\text{DDT-size(f)})/H(p)$, where $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy of $p$.

We may now state our main theorem in the case of functions $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$:

**Theorem 1.1** *Let $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ and let $T$ be a DDT computing $f$. Then*

$$\mathbf{Var}[f] \leq \sum_{i=1}^n \delta_i^{(p)}(T)\, \mathbf{Inf}_i(f).$$

As an immediate corollary we obtain the lower bound on $\mathbf{Inf}_{\max}(f)$ mentioned in Section 1.1:

**Corollary 1.2** *Let $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ with $\Delta^{(p)}(f) \leq d$. Then $\mathbf{Inf}_{\max}(f) \geq \frac{\mathbf{Var}(f)}{d}$.*

**Proof:** Let $T$ be a DDT for $f$ with $\Delta^{(p)}(T) \leq d$. From Theorem 1.1,

$$\mathbf{Var}[f] \leq \sum_{i=1}^n \delta_i^{(p)}(T)\, \mathbf{Inf}_i(f) \leq \mathbf{Inf}_{\max}(f) \sum_{i=1}^n \delta_i^{(p)}(T) = \mathbf{Inf}_{\max}(f) \cdot \Delta^{(p)}(T) \leq \mathbf{Inf}_{\max}(f) \cdot d. \quad \square$$

Some brief comments on our main theorem:

- It is linear in the $\delta_i^{(p)}(T)$'s. Hence if we allow an RDT $\mathcal{T}$ for $f$ and make the natural definition of $\delta_i^{(p)}(\mathcal{T})$, the result still holds by averaging over the distribution $\mathcal{T}$.
- It can be sharp; see Section 3.5 for cases of equality.
- Other corollaries along the lines of Corollary 1.2 follow; for example, if $d$ is an integer then the sum of the influences of the $d$ most influential variables is at least $\mathbf{Var}[f]$.
- In Section 3.3 we will give a "two function" version, which yields an inequality for the case when $T$ is allowed to make a small number of mistakes in computing $f$.

**1.2.1 Influence lower bounds — comparison with previous work.** Proving lower bounds on the influences of boolean functions has had a long history in theoretical computer science, starting with the 1985 paper of Ben-Or and Linial [2] on collective coin flipping. Ben-Or and Linial made the basic observation that if $f : \{-1,1\}^n \to \{-1,1\}$ is balanced, then $\mathbf{Inf}_{\max}(f) \geq \frac{1}{n}$. This follows from the edge isoperimetric inequality on the discrete cube (see, e.g., [6]); however, it is more instructive for us to view it as following from the *Efron-Stein inequality* [10, 27],

$$\mathbf{Var}[f] \leq \mathbf{Inf}(f) = \sum_{i=1}^{n} \mathbf{Inf}_i(f), \tag{1}$$

which holds in the general $p$-biased case, and also in the much more general setting of $f : \Omega \to \mathbb{R}$, where $\Omega$ is a $n$-wise product probability space and $\mathbf{Inf}_i$ is defined appropriately for real-valued functions. (See Appendix A for a more detailed discussion of this setting.) Theorem 1.1 is immediately seen to generalize the Efron-Stein inequality in the case of functions $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$.

Ben-Or and Linial constructed a balanced function $f : \{-1,1\}^n \to \{-1,1\}$ ("Tribes") satisfying $\mathbf{Inf}_{\max}(f) = \Theta(\frac{\log n}{n})$ and conjectured that this was worst possible. There were small improvements on the simple $\frac{1}{n}$ bound ($\frac{2-\epsilon}{n}$ by Alon, $\frac{3-\epsilon}{n}$ by Chor and Gereb-Graus; see [15]) before the famous KKL paper [15] confirmed the conjecture. Note that our theorem improves upon KKL whenever $f$ has $\Delta(f) = o(n/\log n)$; in particular, whenever $f$ has a DDT of size $2^{o(n/\log n)}$.

The KKL result was subsequently generalized by Talagrand [28] who proved that for any $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$,

$$\mathbf{Var}[f] \leq O\Big(\log \frac{1}{p(1-p)}\Big) \sum_{i=1}^{n} \frac{1}{\log_2(8p(1-p)/\mathbf{Inf}_i(f))}\, \mathbf{Inf}_i(f). \tag{2}$$

Talagrand's motivation for proving this was that when $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ is monotone, lower bounds on the sum of $f$'s influences imply "sharp thresholds" for $f$, via the Russo-Margulis lemma [17, 22]. Indeed, this connection with threshold phenomena is one of the chief motivations for studying influences, and it is considered an important problem in the theory of boolean functions and random graphs to provide general conditions under which the total influence is large [7]. Our main inequality provides such a condition: $\mathbf{Inf}(f)$ is large if $f$ has a randomized decision tree $\mathcal{T}$ with $\delta_i^{(p)}(\mathcal{T})$ small for all $i$. Note that when $f$ is a transitive function, this is equivalent to the natural condition that $\Delta^{(p)}(f)$ is small. (See Section 2 for definitions of monotone and transitive functions, as well as further discussion of random graph properties.)

In particular, ours seems to be the first quantitatively strong influence lower bound that takes into account the "structure" or computational complexity of $f$. We note that previously achievable lower bounds on influences in terms of some measure of the complexity of $f$ yield quantitatively much weaker results than can be obtained from our inequality. For instance, Nisan and Szegedy [19] showed that if $f : \{-1,1\}^n \to \{-1,1\}$ is computed by a polynomial over $\mathbb{R}$ of degree $\deg(f)$, then every coordinate $i$ with nonzero influence has $\mathbf{Inf}_i(f) \geq 2^{-\deg(f)}$. Since $D(f) \leq O(\deg(f)^4)$ (by a result of Nisan and Smolensky [8]), our Corollary 1.2 implies that the maximum influence in fact satisfies $\mathbf{Inf}_{\max}(f) \geq \Omega(\mathbf{Var}[f]/\deg(f)^4)$. As another example, suppose $f : \{-1,1\}^n \to \{-1,1\}$ is approximately computed by a polynomial over $\mathbb{R}$ of degree $\widetilde{\deg}(f)$ — i.e. there is a polynomial $p(x)$ of degree $\widetilde{\deg}(f)$ such that $|p(x) - f(x)| < 1/3$ for all $x$. Talagrand's result implies that $\mathbf{Inf}_{\max}(f) \geq \exp(-O(\mathbf{Inf}(f)/\mathbf{Var}[f]))$. Since by [25] we have $\mathbf{Inf}(f) \leq O(\widetilde{\deg}(f))$, one could conclude that $\mathbf{Inf}_{\max}(f) \geq \exp(-O(\widetilde{\deg}(f)/\mathbf{Var}[f]))$. However by contrast, since $D(f) \leq O(\widetilde{\deg}(f)^6)$ by [1], our Corollary 1.2 implies that the maximum influence in fact satisfies $\mathbf{Inf}_{\max}(f) \geq \Omega(\mathbf{Var}[f]/\widetilde{\deg}(f)^6)$.

## 2 Randomized decision tree complexity lower bounds

In this section we give an application of Theorem 1.1 to the problem of randomized decision tree complexity for monotone graph properties. We prove Theorem 1.1 in a more general setting in Section 3.

**2.1 History.** As mentioned in Section 1.1, decision tree complexity has been extensively studied for over three decades. Two special classes of functions have played a prominent role in these investigations. The first is the class of *monotone functions*, those satisfying $f(y) \geq f(x)$ whenever $y \geq x$ under the componentwise partial order. The second is the class of *transitive functions*. An automorphism of the $n$-variate boolean function $f$ is a permutation $\sigma$ of $[n]$ satisfying $f(x_1, \ldots, x_n) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$ for all inputs $x$. We say that $f$ is *transitive* if for each pair $i, j \in [n]$ there is an automorphism of $f$ that sends $i$ to $j$. For example, Rivest and Vuillemin [21] proved that for $n$ a prime power, any $n$-variate monotone transitive function $f$ has $D(f) = n$.

One long studied open question about boolean decision tree complexity is the following: how small can $R(f)$ be in relation to $D(f)$? It is well known [5] that $R(f) \geq \Omega(\sqrt{D(f)})$ for any function $f$, and this is the best general lower bound known. The largest known separation is given by the following recursively defined function: Let $f_0$ be the identity function on a single variable and for $k \geq 1$, let $f_k$ be the function on $n = 4^k$ variables given by $(f^1_{k-1} \wedge f^2_{k-1}) \vee (f^3_{k-1} \wedge f^4_{k-1})$, where $f^i_{k-1}$ is the value of $f_{k-1}$ on the $i$th group of $4^{k-1}$ variables. The function $f_k$ is monotone and transitive, and so by the above result of Rivest and Vuillemin, $D(f_k) = n$. Snir [26] gave an RDT for $f_k$ establishing $R(f) \leq n^\beta$ where $\beta = \log_2\left(\frac{1+\sqrt{33}}{4}\right) \approx 0.753$. Saks and Wigderson [23] proved that Snir's RDT is optimal for $f_k$ and conjectured that $R(f) \geq \Omega(D(f)^\beta)$ for any boolean function; this is not even known to hold for all monotone transitive functions.

A well studied subclass of transitive boolean functions consists of functions derived from graph properties. A *property* of $v$-vertex (undirected) graphs is a set of graphs on vertex set $V = \{1, \ldots, v\}$ that is invariant under vertex relabellings; e.g., the set of graphs on $V$ that are properly 3-colorable. We restrict attention to properties that are non-trivial; i.e., at least one graph has the property and at least one graph does not have the property.

Let $\binom{V}{2}$ denote the set of 2-elements subsets of $V$. Each graph $G$ on $V$ can be identified with the boolean vector $x^G \in \{-1, 1\}^{\binom{V}{2}}$ where $x^G_{\{i,j\}}$ is 1 if $\{i, j\} \in E(G)$ and is $-1$ otherwise. A graph property $\mathcal{P}$ is thus naturally identified with a boolean function $f_{\mathcal{P}} : \{-1, 1\}^{\binom{V}{2}} \longrightarrow \{-1, 1\}$ which maps the vector $x^G$ to 1 if and only if $G$ satisfies $\mathcal{P}$. The invariance of properties under vertex relabellings implies that the associated functions are transitive.

There are examples of graph properties on $v$ vertices that have deterministic decision trees of depth $O(v)$; e.g., the property of being a "scorpion graph" [4]. However, for graph properties that are *monotone* (those whose associated function is monotone), Rivest and Vuillemin [21] proved a lower bound $\Omega(v^2)$ on DDT complexity. A conjecture made by Yao [29] and also attributed to Karp [23] is that this $\Omega(v^2)$ lower bound extends to RDT complexity. This is the problem we make progress on in this section.

Yao observed that an $\Omega(v)$ lower bound for RDT computation of monotone graph properties follows is easy to prove; this also follows from the general bound $R(f) = \Omega(\sqrt{D(f)})$ mentioned earlier. The first improvement on this naive bound came a decade later from Yao himself, who proved an $\Omega(v \log^{1/12} v)$ lower bound using "graph packing" arguments [30]. These arguments were improved by King [16] yielding an $\Omega(v^{5/4})$ lower bound and by Hajnal [14] yielding an $\Omega(v^{4/3})$ lower bound. This lower bound stood for a decade before Chakrabarti and Khot [9] gave a small improvement to $\Omega(v^{4/3} \log^{1/3} v)$. Both the Hajnal and Chakrabarti-Khot bounds have rather long

and technical proofs based on graph packing.

Fairly recently, Friedgut, Kahn and Wigderson [11] proved a general lower bound of a somewhat different form. Given a nonconstant monotone boolean function $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$, it is easy to see that $\mathbf{E}[f]$ is a continuous increasing function of $p$; therefore there is a *critical probability* $p$ for which $\mathbf{E}[f] = 0$, i.e., $\mathbf{Var}[f] = 1$. Friedgut, Kahn and Wigderson proved that any nontrivial monotone $v$-vertex graph property has RDT complexity $\Omega(\min\{\frac{v}{\min(p,1-p)}, \frac{v^2}{\log v}\})$ when $p$ is the critical probability for $f$. In fact they show that $\Delta^{(p)}(f)$ is at least this quantity. The FKW bound can improve on Chakrabarthi-Khot in cases where the critical probability is sufficiently close to 0 or 1. We remark that the proof in FKW also uses a graph packing argument.

**2.2 Our result.** As a simple consequence of our elementary main inequality Theorem 1.1 and a recent elementary inequality from [20], we obtain the following:

**Theorem 2.1** *Let $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ be a nonconstant monotone transitive function, where $p$ is the critical probability for $f$. Write $q = 1 - p$. Then*

$$R(f) \geq \Delta^{(p)}(f) \geq \frac{n^{2/3}}{(4pq)^{1/3}}.$$

*In particular,*

$$R(f) \geq \Delta^{(p)}(f) = \frac{v^{4/3}}{(32pq)^{1/3}}$$

*if $f$ corresponds to a $v$-vertex graph property.*

**Proof:** The inequality we need from [20] is the following:

$$\text{For all } p, \text{ if } f : \{-1,1\}^n_{(p)} \to \{-1,1\} \text{ is monotone then } \mathbf{Inf}(f) \leq 2\sqrt{pq}\sqrt{\Delta^{(p)}(f)}. \tag{3}$$

Fix $p$ to be the critical probability of $f$ and let $T$ be the DDT computing $f$ with expected cost $\Delta^{(p)}(f)$. We apply Theorem 1.1, using $\mathbf{Var}[f] = 1$ since $p$ is critical and $\mathbf{Inf}_i(f) = \mathbf{Inf}(f)/n$ since $f$ is transitive (and hence all coordinates have the same influence). This gives $1 \leq (\mathbf{Inf}(f)/n) \cdot \Delta^{(p)}(f)$. Using (3) to bound $\mathbf{Inf}(f)$ we get $1 \leq (2\sqrt{pq}/n) \cdot (\Delta^{(p)}(f))^{3/2}$, and this can be rearranged to give the desired result. □

**2.3 Discussion.** In the case of monotone graph properties, our result always improves on Hajnal's $\Omega(v^{4/3})$ lower bound and can be superior to both Chakrabarti-Khot (when $\min\{p,q\}$ is small enough) and to FKW (when $\min\{p,q\}$ is large enough). It is worth noting that unlike all previous lower bounds for monotone graph properties, our proof makes no use of graph packing arguments, instead relying only on elementary probabilistic arguments (see Appendix A for a generalized proof of [20]'s inequality, noting that the proof is even simpler when specialized back to (3)).

Most interestingly, we obtain a result essentially as good as the best unconditional bound (Chakrabarti-Khot) in the more general context of monotone *transitive* functions, not just graph properties. Further, our bound for monotone transitive functions is known to be essentially tight in the case of $p = \frac{1}{2}$: in [3], a family $(f_n)$ of $\frac{1}{2}$-critical monotone transitive functions is presented with $\Delta(f_n) \leq O(n^{2/3} \log n)$. It's quite curious to note that the place where the RDT complexity of monotone graph properties has been stuck for almost 15 years, $v^{4/3}$, is exactly the tight bound for monotone transitive functions. Perhaps this suggests that in some way the argument of Hajnal is not really using the fact that $f$ is a graph property — just that it's transitive. Indeed, one might wonder the same thing about Chakrabarti-Khot, since their $v^{4/3} \log^{1/3} v$ lower bound could also hold for monotone transitive functions — the example of [3] is not able to rule it out.

## 3 The main inequality

**3.1 Decision trees, variation, influences — general definitions.** The proof of Theorem 1.1 is most naturally carried out in a significantly more general context than that of functions $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$. Specifically, we will consider functions

$$f : \Omega \longrightarrow Z$$

mapping a *product probability space* into a *metric space*. In this section we give the necessary definitions.

Let us begin with the domain. Here we have an $n$-wise product probability space $\Omega = (X, \mu)$, meaning that that the underlying set $X$ is a product set $X_1 \times \cdots \times X_n$ and the measure $\mu$ is a product probability measure $\mu_1 \times \cdots \times \mu_n$, where $\mu_i$ is a measure on $X_i$.[1] We write $\Omega_i$ for the probability space $(X_i, \mu_i)$. We use the notation $x \leftarrow \Omega$ to mean that $x$ is an element of $X$ randomly selected according to $\Omega$.

The range of our functions is a metric space $(Z, d)$. (Actually we can allow a "pseudo-metric", meaning we relax the condition that $d(z, z') = 0 \Rightarrow z = z'$.) Useful examples to keep in mind are the following: $Z$ any finite set with $d(z, z') = \mathbf{1}_{z=z'}$; and, $Z = \mathbb{R}$ with $d(z, z') = |z - z'|$. Of course, in the special case of boolean-valued functions, $Z = \{-1, 1\}$, all metrics are the same up to a constant factor.

We now give the definitions of decision trees, variation, and influences for functions $f : \Omega \to Z$.

The definitions of decision trees in the context of functions mapping a product set domain $X = X_1 \times \cdots \times X_n$ into a set $Z$ are the obvious ones. Briefly, a DDT will be a rooted directed tree $T$ in which each internal node $v$ is labelled by a coordinate $i_v \in [n]$ and each leaf is labelled by an element of the output set $Z$. Further, the arcs emanating from each internal node $v$ must be in one-to-one correspondence with $X_{i_v}$. The node labels along every root-leaf path are required to be distinct. $T$ computes a function $f_T : X \to Z$ in the obvious way; we retain the notion of the cost of $T$ on input $x$ as the length of the root-leaf path $T$ follows on input $x$. Thus we have the usual notions of $D(T)$ and $D(f)$, and also the (zero-error) randomized decision tree complexities $R(\mathcal{T})$ and $R(f)$. With the product probability measure $\mu$ on $X$, we can also naturally extend our notions of *expected cost* from Section 1.2: given a DDT $T$ computing $f$,

$$\delta_i^\mu(T) = \mathop{\mathbf{Pr}}_{x \leftarrow \Omega = (X,\mu)}[T \text{ queries } x_i],$$

and $\Delta^\mu(T)$ and $\Delta^\mu(f)$ are similarly defined. We will henceforth drop the superscript $\mu$ when it is clear from context. Note that as before we have $\Delta(f) \le R(f)$ (assuming, without loss of generality, that $\mu$'s support is all of $X$).

We now give the definitions of *variation* and *influences* for functions $f : \Omega \to Z$. The *variation* of $f : \Omega \to Z$ is

$$\mathbf{Vr}^{\mu,d}[f] = \mathop{\mathbf{E}}_{(x,y) \leftarrow \Omega \times \Omega}[d(f(x), f(y))].$$

To define influences, first let $\Omega^{(i)}$ denote the probability space given by pairs $(x, x^{(i)})$, where $x$ is chosen from $\Omega$ and $x^{(i)}$ is formed by rerandomizing the $i$th coordinate of $x$ using $\mu_i$. Then the *influence of the $i$th coordinate* on $f : \Omega \to Z$ is defined to be

$$\mathbf{Inf}_i^{\mu,d}(f) = \mathop{\mathbf{E}}_{(x,x^{(i)}) \leftarrow \Omega^{(i)}}[d(f(x), f(x^{(i)}))].$$

[1] For simplicity, we assume in this paper that $X$ is finite.

We will usually drop the superscripts $\mu$ and $d$ on $\mathbf{Vr}$ and $\mathbf{Inf}_i$ when they are implied by context. Note that if we view the functions $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ from Section 1 as mapping into the metric space on $\{-1,1\}$ with distance $d$ given by $d(z,z') = |z-z'| = 2\cdot\mathbf{1}_{z\neq z'}$, then we get agreement in the definitions of $\mathbf{Inf}_i(f)$ and also $\mathbf{Vr}[f] = \mathbf{Var}[f]$.

**3.2 Theorem and proof.** We now state and prove our main inequality, which includes Theorem 1.1 as a special case.

**Theorem 3.1** *Let $f : \Omega \to (Z,d)$ be a function mapping an $n$-wise product probability space into a metric space, and let $T$ be a DDT computing $f$. Then*

$$\mathbf{Vr}[f] \leq \sum_{i=1}^{n} \delta_i(T)\, \mathbf{Inf}_i(f).$$

**Proof:** Let $x$ and $y$ be random inputs chosen independently from $\Omega$. Given a subset $J \subseteq [n]$ we will write $x_J y$ for the *hybrid input* in $X$ that agrees with $x$ on the coordinates in $J$ and with $y$ on the coordinates in $[n] \setminus J$. Let $i_1, \ldots, i_d$ denote the sequence of variables queried by $T$ on input $x$ (these $i$'s are random variables and $d$ is also a random variable). For $t \geq 0$, let $J[t] = \{i_r : r > t\}$. Finally, let $u[t] = x_{J[t]}y$. All $\mathbf{E}[\cdot]$'s and $\mathbf{Pr}[\cdot]$'s in what follows are over all the random variables just described (i.e., $x$, $y$, $i$'s, $d$, $u[\cdot]$'s).

We begin with the simple observation

$$\mathbf{Vr}[f] = \mathbf{E}[d(f(x), f(y))] = \mathbf{E}[d(f(u[0]), f(u[d]))],$$

which follows because $y = u[d]$ and $f(x) = f(u[0])$ (although $x$ does not necessarily equal $u[0]$). This latter equality is the only place in the proof we use the fact that $T$ computes $f$.

We next make the obvious step

$$\mathbf{E}[d(f(u[0]), f(u[d]))] \leq \mathbf{E}\Big[\sum_{t\geq 1} d(f(u[t-1]), f(u[t]))\Big]$$

which uses the fact that $d$ is a metric (and the fact that $u[t] = u[d]$ for all $t \geq d$). We now use linearity of expectation and then condition on the value of $i_t$:

$$\mathbf{E}\Big[\sum_{t\geq 1} d(f(u[t-1]), f(u[t]))\Big] = \sum_{t\geq 1}\sum_{i=1}^{n} \mathbf{Pr}[i_t = i]\; \mathbf{E}\Big[d(f(u[t-1]), f(u[t])) \;\Big|\; i_t = i\Big]. \tag{4}$$

(Technically, $i_t$ may be "undefined" when $t > d$, but this doesn't matter since $u[t] = u[t+1]$ then and hence 0 is contributed to the expectation.)

We now come to the key observation of the proof: we claim that conditioned on $i_t = i$, $(u[t-1], u[t])$ is distributed according to $\Omega^{(i)}$. Certainly conditioning on $i_t = i$ imposes some constraints on $x_{i_0}, \ldots, x_{i_{t-1}}$. However it is clear that $x_{i_t}, \ldots, x_{i_d}$ are independent of both $i_t$ and $x_{i_0}, \ldots, x_{i_{t-1}}$, and this is of course also true of all coordinates of $y$. Since $x_{i_0}, \ldots, x_{i_{t-1}}$ are rerandomized using $y$'s values in the formation of $u[t-1]$, we conclude that $u[t-1]$ conditioned on $i_t = i$ is simply distributed according to $\Omega$. And then conditioned on $i_t = i$, $u[t]$ is distributed as $u[t-1]$ with the $i$th coordinate rerandomized. Hence we have justified the claim that, conditioned on $i_t = i$, $(u[t-1], u[t])$ is distributed according to $\Omega^{(i)}$.

Thus we have

$$(4) = \sum_{t\geq 1}\sum_{i=1}^{n} \mathbf{Pr}[i_t = i]\; \mathbf{Inf}_{i_t}(f) = \sum_{i=1}^{n}\Big(\sum_{t\geq 0} \mathbf{Pr}[i_t = i]\Big)\; \mathbf{Inf}_i(f) = \sum_{i=1}^{n} \delta_i(T)\; \mathbf{Inf}_i(f),$$

and the proof is complete. □

**3.3 Corollaries and two function version.** In this section we treat some immediate corollaries of Theorem 3.1. Certainly the analogue of Corollary 1.2 holds for Theorem 3.1, as do the first and third remarks made at the end of Section 1.2. We now give the promised "two function" version. Define

$$\mathbf{CoVr}[f,g] = \mathop{\mathbf{E}}_{(x,y)\leftarrow\Omega\times\Omega}[d(f(x),g(y))] - \mathop{\mathbf{E}}_{x\leftarrow\Omega}[d(f(x),g(x))],$$

so in particular $\mathbf{CoVr}[f,f] = \mathbf{Vr}[f]$. Thus the following theorem generalizes Theorem 3.1:

**Theorem 3.2** *Let $f, g : \Omega \to (Z,d)$ be functions mapping an $n$-wise product probability space into a metric space, and let $\mathcal{T}$ be an RDT computing $f$. Then*

$$\left|\mathbf{CoVr}[f,g]\right| \leq \sum_{i=1}^{n} \delta_i(\mathcal{T})\ \mathbf{Inf}_i(g).$$

**Proof:** As usual we can assume by averaging that $\mathcal{T}$ is a DDT $T$ computing $f$. Using the same setup as in the proof of Theorem 3.1, we have

$$\mathbf{CoVr}[f,g] = \mathbf{E}[d(f(x),g(y))] - \mathbf{E}[d(f(u[0]),g(u[0]))] = \mathbf{E}[d(f(u[0]),g(u[d]))] - \mathbf{E}[d(f(u[0]),g(u[0]))]$$

where in the first equality we used that $u[0]$ is, in isolation, distributed according to $\Omega$, and in the second equality we used the fact that $f(x) = f(u[0])$ since $T$ computes $f$ (as in the previous proof). Now using the fact that $d$ is a metric we get

$$\mathbf{CoVr}[f,g] = \mathbf{E}[d(f(u[0]),g(u[d]))] - \mathbf{E}[d(f(u[0]),g(u[0]))] \leq \mathbf{E}[d(g(u[0]),g(u[d]))]$$

and of course this is also true for $-\mathbf{CoVr}[f,g]$. The proof now proceeds exactly as before with $g$ in place of $f$; note that from this point on in the previous proof we did not use the fact that $T$ computed $f$. □

We give an alternate two function extension in Appendix B.

As mentioned at the end of Section 1.2, Theorem 3.2 can be useful when $g$ is a function that is "close" to having a good decision tree. Consider, for example, the case when $Z$ is an arbitrary set and $d(z,z') = \mathbf{1}_{z\neq z'}$. Now suppose $g$ is $\epsilon$-close to some function $f : \Omega \to (Z,d)$ having a good randomized decision tree $\mathcal{T}$, meaning that $\mathbf{Pr}[g \neq f] = \mathbf{E}[d(f(x),g(x))] = \epsilon$. Write $Z = \{z_1,\ldots,z_s\}$ and assume without loss of generality that $p_1 \leq p_2 \leq \cdots \leq p_s$, where $p_i = \mathbf{Pr}[g(x) = z_i]$. Let us further assume that $\epsilon$ is "small"; specifically, that $\epsilon \leq p_1$.

We have $\mathbf{CoVr}[f,g] = \mathbf{Pr}[f(x) \neq g(y)] - \epsilon$, where $x$ and $y$ are independent, and $\mathbf{Vr}[g] = \sum_{i=1}^{s} p_i(1-p_i)$. It's not too hard to see that, subject to $\mathbf{Pr}[f \neq g] = \epsilon$, we have that $\mathbf{Pr}[f(x) \neq g(y)]$ is minimized when the following holds: $f(x) = g(x)$ whenever $g(x) \neq z_1$; further, when $g(x) = z_1$, $f(x) = z_1$ with probability $1 - \epsilon/p_1$ and is $z_s$ otherwise. In this case,

$$\mathbf{Pr}[f(x) \neq g(y)] = p_1(1-p_1+\epsilon) + p_s(1-p_s-\epsilon) + \sum_{i=2}^{s-1} p_i(1-p_i) = \mathbf{Vr}[g] - (p_s - p_1)\epsilon,$$

and hence we conclude

$$\mathbf{Vr}[g] - 2\epsilon \leq \mathbf{Vr}[g] - (1 + (p_s - p_1))\epsilon = \mathbf{CoVr}[f,g] \leq \sum_{i=1}^{n} \delta_i(\mathcal{T})\ \mathbf{Inf}_i(g),$$

which can provide a good lower bound on $\mathbf{Inf}_{\max}(g)$ if $\epsilon$ is small and the $\delta_i$'s for $f$ are small.

**3.4 When $d$ is not a metric.** In this section we generalize our results to the case when $f$ maps into $(Z, \rho)$, where $(Z, \rho)$ is a "semimetric". This just means that $\rho$ need not satisfy the triangle inequality; specifically, all we require of $\rho$ is that $\rho \geq 0$, $\rho(z, z) = 0$, and $\rho(z, z') = \rho(z', z)$. (Again we do not insist that $\rho(z, z') = 0 \Rightarrow z = z'$.) Our main motivation for studying this extension is the case $Z = \mathbb{R}$ with $\rho = \rho_2(z, z') := (z - z')^2/2$. In this case $\mathbf{Vr}^{\rho_2}[f] = \mathbf{Var}[f]$ and $\mathbf{Inf}^{\rho_2}(f)$ has the meaning commonly associated with this notation for functions $f : \Omega \to \mathbb{R}$; e.g., the interpretation used in the Efron-Stein inequality. See Appendix A for further discussion.

To study the semimetric case, we simply introduce a quantity measuring the extent to which the triangle inequality fails for $\rho$ on paths of length $k$. We define the *defect* of a sequence $z_0, z_1, \ldots, z_k \in Z^{k+1}$ to be $\rho(z_0, z_k) \;/\; (\sum_{t=1}^{k} \rho(z_{t-1}, z_t))$, where $\frac{0}{0}$ is taken to be 1. We then define the *$k$-defect of* $\rho$, denoted $\mathbf{Def}_k(\rho)$, to be the maximum defect of any sequence $z_0, \ldots, z_k$. The following facts are easy to check:

- $\mathbf{Def}_1(\rho) = 1$ and $\mathbf{Def}_k(\rho)$ is nondecreasing with $k$.
- $\mathbf{Def}_k(\rho) \leq (\sup \rho)/(\inf \rho)$ for all $k$.
- $\mathbf{Def}_2(\rho) = 1$ implies that $\rho$ satisfies the triangle inequality, which, in turn, implies that $\mathbf{Def}_k(\rho) = 1$ for all $k$; i.e., $\rho$ is a metric.
- If $\rho^{1/q}$ is a metric for some $q \geq 1$ then $\mathbf{Def}_k(\rho) \leq k^{q-1}$. Thus in our motivating case with $Z = \mathbb{R}$ and $\rho(z, z') = (z - z')^2/2$ we have $\mathbf{Def}_k(\rho) \leq k$.
- If $Z \subseteq \mathbb{R}$ and $\rho(z, z') = |z - z'|^q$ for some $q \geq 1$, then $\mathbf{Def}_k(\rho) \leq |Z|^{q-1}$ for all $k$.

It is easy to see how to generalize Theorems 3.1 and 3.2 for semimetrics $\rho$; since Theorem 3.2 is more general, we will only state its extension:

**Theorem 3.3** *Let $f, g : \Omega \to (Z, \rho)$ be functions mapping an $n$-wise product probability space into a semimetric space, and let $\mathcal{T}$ be an RDT computing $f$. Let $k$ be the length of the longest path in any DDT in $\mathcal{T}$'s support. Then*

$$\left|\mathbf{CoVr}[f, g]\right| \leq \mathbf{Def}_k(\rho) \sum_{i=1}^{n} \delta_i(\mathcal{T})\, \mathbf{Inf}_i(g).$$

This is the most general version of our main inequality that we prove. In the semimetric setting we are most interested in, namely that of one function $f : \Omega \to (\mathbb{R}, \rho_2)$, we have the following:

**Corollary 3.4** *Let $f : \Omega \to (\mathbb{R}, \rho_2)$ be a function mapping an $n$-wise product probability space into the real line with semimetric $\rho_2(z, z') = (z - z')^2/2$, and let $\mathcal{T}$ be an RDT computing $f$. Let $k$ be the length of the longest path in any DDT in $\mathcal{T}$'s support. Then*

$$\mathbf{Var}[f] \leq k \sum_{i=1}^{n} \delta_i(\mathcal{T})\, \mathbf{Inf}_i^{\rho_2}(f),$$

*In particular, if $D(f) \leq k$ then $f$ must have a coordinate with $\rho_2$-influence at least $\mathbf{Var}[f]/k^2$.*

**3.5 Tightness of the inequality** Our main Theorem 3.1 can be tight; one class of DDTs for which it is tight are *read-once* decision trees. These are simply decision trees in which each internal node queries a different variable. It is quite easy to see that equality is achieved in Theorem 3.1 when $T$ is read-once. The only inequality in the proof comes from

$$d(f(u[0]), f(u[d])) \leq \sum_{t \geq 1} d(f(u[t-1]), f(u[t])). \tag{5}$$

If $T$ is read-once then the first time $u[t-1]$ and $u[t]$ differ, further replacement of the variables $i_{t+1}$ cannot affect the path $T$ follows, and hence $f(u[t]) = f(u[t+1]) = f(u[t+2]) = \cdots$. Thus the sequence $f(u[0]), f(u[1]), \ldots, f(u[d])$ changes value at most once, hence (5) and thus Theorem 3.1 have equality. Note that this argument shows that equality holds even if $d = \rho$ is just a semimetric.

Simple examples of read-once DDTs are those for AND : $\{-1,1\}^n \to \{-1,1\}$ and OR : $\{-1,1\}^n \to \{-1,1\}$. The simplest nontrivial balanced example is the "selection function" SEL : $\{-1,1\}^3 \to \{-1,1\}$ which maps $(x_1, x_2, x_3)$ to $x_2$ if $x_1 = 1$, or $x_3$ if $x_1 = -1$.

To get a more general family of examples we introduce the notion of "recursively read-once trees". The definition is as follows: Any read-once DDT is recursively read-once. Now suppose that $T$ is a recursively read-once DDT computing the function $f_T : X_1 \times \cdots \times X_n \to Z$ and that $U_1, \ldots, U_n$ are also recursively read-once DDTs, where $U_i$ computes the function $f_{U_i} : Y_1^i \times \cdots \times Y_{m_i}^i \to X_i$. Then we have a composition function $F = f_T(f_{U_1}, \ldots, f_{U_n})$ on the product domain $Y_1^1 \times Y_2^1 \times \cdots \times Y_{m_n}^n$. Further, there is a natural DDT $V = T(U_1, \ldots, U_n)$ computing $F$, and we stipulate that this DDT is also recursively read-once. Note that a recursively read-once DDT will not, in general, be read-once.

It is not hard to check that recursively read-once trees compute functions $f$ that are tight for Theorem 3.1; by induction one can show that the sequences $f(u[0]), f(u[1]), \ldots, f(u[d])$ arising in the proof still have the property that they change value at most once. Hence Theorem 3.1 is tight for any Tribes-type function (OR of disjoint ANDs) and the natural (non-read-once) DDT computing it; as another example, it is tight for the recursive SEL function $\mathrm{SEL}(\mathrm{SEL}(\cdots), \mathrm{SEL}(\cdots), \mathrm{SEL}(\cdots))$ of any depth and the natural (non-read-once) DDT computing it.

Finally, we discuss the necessity of the factor $\mathbf{Def}(\rho)$ in the "one function" version of Theorem 3.3. We do not have any general family of examples showing the necessity of this factor. Indeed, as far as we know, it may be possible to replace the factor $\mathbf{Def}(\rho)$ by an absolute constant. This possibility is particularly intriguing in the case of $\rho = \rho_2$ from Corollary 3.4 ([24] raised a similar question which turns out to be related — see Appendix A). However we *can* show that the inequality of Theorem 3.3 with the constant 1 in place of $\mathbf{Def}(\rho)$ does not hold, even in the $(\mathbb{R}, \rho_2)$ case. The $\{-1,1\}^3 \to (\mathbb{R}, \rho_2)$ example shown in Figure 1 — naturally, not a read-once tree — demonstrates that a constant slightly greater than 1 is necessary. Except for optimizing the leaf labels in this particular tree, this is the worst example we know.

## 4 Questions for Future Work

- Is it possible to explain the "coincidence" that our near-tight lower bound on $\Delta^{(p)}(f)$ for monotone transitive functions gives a lower bound for graph properties — about $v^{4/3}$ — that essentially matches the lower bound barrier that has stood since Hajnal '91? Perhaps either the Hajnal/Chakrabarti-Khot arguments can be reframed in terms of merely transitive functions (if true of Chakrabarti-Khot, this would be quite interesting); or, perhaps graph-theoretic arguments can augment our elementary probabilistic reasoning to produce a better lower bound.

- Can our inequality in the real-valued, $\rho_2$ case — Corollary 3.4 — be sharpened? If the factor

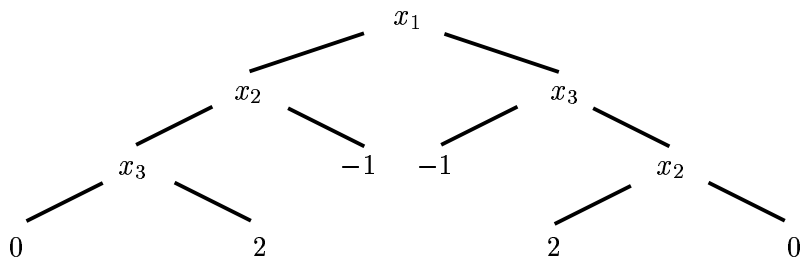


Figure 1: Left edges correspond to $-1$, right edges to 1. The function $f : \{-1,1\}^3 \to \mathbb{R}$ computed by $T$ has $\mathbf{Var}[f] = \frac{3}{2}$, but $(\delta_1(T), \delta_2(T), \delta_3(T)) = (1, \frac{3}{4}, \frac{3}{4})$ and $(\mathbf{Inf}_1^{\rho_2}(f), \mathbf{Inf}_2^{\rho_2}(f), \mathbf{Inf}_3^{\rho_2}(f)) = (\frac{1}{8}, \frac{7}{8}, \frac{7}{8})$, where $\rho_2(x,y) = (x-y)^2/2$, so $\sum_{i=1}^3 \delta_i(T)\, \mathbf{Inf}_i^{\rho_2}(f) = \frac{23}{16} < \frac{3}{2}$.

$k$ could be replaced by a universal constant, this would be a very strong near-sharpening of the Efron-Stein inequality.

- What other applications might our main inequality have? We suggest there might be applications in computational learning theory or in the theory of random graphs.

## 5 Acknowledgments

We would like to thank Andris Ambainis, Laci Lovasz, and Avi Wigderson for helpful discussions.

## References

[1] R. Beals, H. Buhrman, R. Cleve, M. Mosca, and R. de Wolf. Quantum lower bounds by polynomials. *Journal of the ACM*, 48(4):778–797, 2001.

[2] M. Ben-Or and N. Linial. Collective coin flipping. In *Proceedings of the 26th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 408–416, 1985.

[3] I. Benjamini, O. Schramm, and D. Wilson. Balanced Boolean functions that can be evaluated so that every input bit is unlikely to be read. To appear in *STOC* 2005, 2005.

[4] M. Best, P. van Emde Boas, and H. Lenstra. A sharpened version of the aanderaa-rosenberg conjecture. Technical Report Report ZW30/74, Mathematisch Centrum Amsterdam, 1974.

[5] M. Blum and R. Impagliazzo. Generic oracles and oracle classes. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, pages 118–126, 1987.

[6] B. Bollobas. *Combinatorics: Set Systems, Hypergraphs, Families of Vectors and Combinatorial Probability.* Cambridge University Press, 1986.

[7] J. Bourgain and G. Kalai. Influences of variables and threshold intervals under group symmetries. *GAFA*, 7:438–461, 1997.

[8] H. Buhrman and R. de Wolf. Complexity measures and decision tree complexity: a survey. *Theoretical Computer Science*, 288(1):21–43, 2002.

[9] A. Chakrabarti and S. Khot. Improved lower bounds on the randomized complexity of graph properties. In *Proceedings of the 28th International Colloquium on Automata, Languages and Programming*, pages 285–296, 2001.

[10] B. Efron and C. Stein. The jackknife estimate of variance. *Annals of Statistics*, 9:586–596, 1981.

[11] E. Friedgut, J. Kahn, and A. Wigderson. Computing graph properties by randomized subcube partitions. In *Proceedings of the 6th International Workshop on Random and Approximation Techniques*, pages 105–113, 2002.

[12] E. Friedgut and G. Kalai. Every Monotone Graph Property has a Sharp Threshold. *Proceedings of the AMS*, 124:2993–3002, 1996.

[13] L. Gross. Logarithmic Sobolev inequalities. *Amer. J. Math.*, 97(4):1061–1083, 1975.

[14] A. Hajnal. An $\Omega(n^{4/3})$ lower bound on the randomized complexity of graph properties. *Combinatorica*, 11:131–143, 1991.

[15] J. Kahn, G. Kalai, and N. Linial. The influence of variables on Boolean functions. In *Proceedings of the 29th Annual Symposium on Foundations of Computer Science*, pages 68–80, 1988.

[16] V. King. Lower bounds on the complexity of graph properties. In *Proceedings of the 20th Annual Symposium on Theory of Computing*, pages 468–476, 1988.

[17] G. Margulis. Probabilistic characteristics of graphs with large connectivity. *Prob. Peredachi Inform.*, 10:101–108, 1974.

[18] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. Manuscript, 2005.

[19] N. Nisan and M. Szegedy. On the degree of Boolean functions as real polynomials. In *Proceedings of the Twenty-Fourth Annual Symposium on Theory of Computing*, pages 462–467, 1992.

[20] R. O'Donnell and R. Servedio. Learning monotone functions from random examples in polynomial time. manuscript, 2005.

[21] R. Rivest and J. Vuillemin. On recognizing graph properties from adjacency matrices. *Theoretical Computer Science*, 3:371–384, 1976.

[22] L. Russo. On the critical percolation probabilities. *Z. Wahrsch. werw. Gebiete*, 43:39–48, 1978.

[23] M. Saks and A. Wigderson. Probabilistic boolean decision trees and the complexity of evaluating game trees. In *Proceedings of the 27th Annual Symposium on Foundations of Computer Science*, pages 29–38, 1986.

[24] O. Schramm and J. Steif. Quantitative noise sensitivity and exceptional times for percolation. Manuscript, 2005.

[25] Y. Shi. Lower bounds of quantum black-box complexity and degree of approximating polynomials by influence of boolean variables. *Information Processing Letters*, 75(1-2):79–83, 2000.

[26] M. Snir. Lower bounds for probabilistic linear decision trees. *Theoretical Computer Science*, 38:69–82, 1985.

[27] J. M. Steele. An Efron-Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14:753–758, 1986.

[28] M. Talagrand. On russo's approximate 0-1 law. *The Annals of Probability*, 22(3):1576–1587, 1994.

[29] A. Yao. Probabilistic computations: Towards a unified measure of complexity. In *Proceedings of the Seventeenth Annual Symposium on Foundations of Computer Science*, pages 222–227, 1977.

[30] A. Yao. Lower bounds to randomized algorithms for graph properties. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, pages 393–400, 1987.

## A Real-valued functions

In this section we discuss further the case of functions $f : \Omega \to (\mathbb{R}, \rho_2)$, where $\Omega$ is an $n$-wise probability space and $\rho_2$ is the semimetric $\rho_2(z, z') = (z - z')^2/2$. As noted in Section 3.4, in this case we have $\mathbf{Vr}^{\rho_2}[f] = \mathbf{Var}[f]$, the usual variance of $f$; also

$$\mathbf{Inf}_i^{\rho_2}[f] = \mathbf{E}_\Omega[\mathbf{Var}_{\Omega_i}[f]].$$

For this section only, we will drop the superscript $\rho_2$ on $\mathbf{Inf}_i$. This quadratic, real-valued notion of "influence" arises naturally in the Efron-Stein inequality [10, 27] and other discrete log-Sobolev and Poincaré inequalities [13], in Talgrand's result (2) from [28], and elsewhere [18]. Efron-Stein (1) and Talgrand's (2) both hold with this definition of $\mathbf{Inf}_i(f)$.

Recall our inequality in this setting, Corollary 3.4:

$$\mathbf{Var}[f] \le k \sum_{i=1}^{n} \delta_i(\mathcal{T})\, \mathbf{Inf}_i(f),$$

where $\mathcal{T}$ is an RDT computing $f$ and $k$ is the maximum path length in any DDT in $\mathcal{T}$. This inequality is incomparable with Efron-Stein, but it is tempting to wonder if the factor $k$ can be replaced with an absolute constant; in this case, we would, like Talagrand, get an inequality strictly better than Efron-Stein "up to constants". However, as shown in Section 3.5, the factor $k$ cannot be replaced with the constant 1.

In the remainder of this section, we will give a slight sharpening of Corollary 3.4 via a completely different method of proof. Intriguingly, the proof is by giving a common generalization of the proofs of two other recent inequalities. The first is the main inequality from [20] (a paper on learning monotone functions), a special case of which (3) we used in the proof of Theorem 2.1: If $f : \{-1, 1\}^n_{(p)} \to \mathbb{R}$ is computed by the randomized decision tree $\mathcal{T}$, then

$$\sum_{i=1}^{n} \hat{f}(\{i\}) \le \|f\|_2 \sqrt{\Delta^{(p)}(\mathcal{T})}. \tag{6}$$

Here $\hat{f}$ denotes the $p$-biased Fourier transform of $f$; we will discuss Fourier transforms shortly. The second inequality generalized is the main technical inequality from [24] (a paper on exceptional times for percolation problems): If $f : \{-1, 1\}^n \to \mathbb{R}$ is computed by the randomized decision tree $\mathcal{T}$, then for all $k = 1 \ldots n$,

$$\sum_{|S|=k} \hat{f}(S)^2 \le (\max \delta_i(\mathcal{T})) \cdot k \cdot \|f\|_2^2. \tag{7}$$

It should be noted that (6) was proved in the more general setting that $\mathcal{T}$ is a randomized *subcube partition* computing $f$ (see [20] for the appropriate definition). We note that our generalizing

inequality, Theorem A.1 below, is also easily seen to hold for randomized subcube partitions; however, we omit this extension for simplicity.

To state our generalized inequality for functions $f : \Omega \to \mathbb{R}$ we need to use an orthogonal decomposition of $\Omega$ and the associated "Fourier transform" of $f$. (The same is used in some proofs of the Efron-Stein inequality.) The space of functions $\Omega \to \mathbb{R}$ is an inner product space under the inner product $\langle f, g \rangle = \mathbf{E}_{x \in \Omega}[f(x)g(x)]$. Given $x \in X$ and $S \subseteq [n]$, let $x_S$ denote $\{x_i : i \in S\}$. Now $\Omega$ is an orthogonal sum of spaces $\Omega = \oplus_{S \subseteq [n]} \Omega_S$, where $\Omega_S$ denotes the space of all functions $f : \Omega \to \mathbb{R}$ such that

- $f(x)$ depends only on $x_S$, and
- $f$ is orthogonal to all functions in the spaces $X_{S'}$ for $S' \subsetneq S$.

Hence we can write any $f : \Omega \to \mathbb{R}$ as

$$f(x) = \sum_{S \subseteq [n]} f_S(x), \tag{8}$$

where $f_S$ is the projection of $f$ onto $\Omega_S$. We will refer to (8) as the "Fourier expansion" of $f$, since it agrees with the usual Fourier expansions for $\{-1,1\}^n$ and $\{-1,1\}^n_{(p)}$ when $f_S(x)$ is identified with $\hat{f}(S)\chi_S(x)$. It is straightforward to verify "Plancherel's identity"

$$\langle f, g \rangle = \sum_{S \subseteq [n]} \langle f_S, g_S \rangle,$$

as well as the formulas for variance and influences:

$$\mathbf{Var}[f] = \sum_{S \neq \emptyset} \|f_S\|_2^2, \qquad \mathbf{Inf}_i(f) = \sum_{S \ni i} \|f_S\|_2^2.$$

Summing the influence formulas we get

$$\mathbf{Inf}(f) = \sum_{i=1}^{n} \mathbf{Inf}_i(f) = \sum_{S \subseteq [n]} |S| \cdot \|f_S\|_2^2.$$

Incidentally, with these formulas in hand the Efron-Stein inequality $\mathbf{Var}[f] \leq \mathbf{Inf}(f)$ becomes immediate.

We can now give our generalized inequality. Recall that an *antichain* on $[n]$ is a family of subsets $\mathcal{S}$ such that $S \subsetneq S'$ never holds for $S, S' \in \mathcal{S}$. For functions $f, g$, we denote their covariance by

$$\mathbf{Cov}[f, g] = \mathbf{E}[fg] - \mathbf{E}[f]\,\mathbf{E}[g]\,.$$

**Theorem A.1** *Let $f : \Omega \to \mathbb{R}$ be a real-valued function on an $n$-wise product probability space, and let $\mathcal{T}$ be an RDT computing $f$. Let $g : \Omega \to \mathbb{R}$ be a function whose Fourier expansion is supported on an antichain; i.e., for some antichain $\mathcal{S}$ of $[n]$, it holds that $g_S = 0$ for all $S \notin \mathcal{S}$. Then*

$$\mathbf{Cov}[f, g]^2 \leq \mathbf{Var}[f] \cdot \sum_{i=1}^{n} \delta_i(\mathcal{T})\, \mathbf{Inf}_i(g).$$

Before proving Theorem A.1, let us see how it implies inequalities (6), (7), and Corollary 3.4. For (6) we take $g = \sum_{i=1}^{n} \chi_{\{i\}}$; we then get

$$\left(\sum_{i=1}^{n} \hat{f}(\{i\})\right)^2 \leq \mathbf{Var}[f] \cdot \sum_{i=1}^{n} \delta_i(\mathcal{T}).$$

After taking a square root, this is equivalent to (6) (since $f$ may be replaced by $f - \mathbf{E}[f]$). For (7) we take $g = \sum_{|S|=k} \hat{f}(S)\chi_S$ and obtain

$$\begin{aligned}\left(\sum_{|S|=k} \hat{f}(S)^2\right)^2 &\leq \mathbf{Var}[f] \cdot \sum_{i=1}^{n} \delta_i(\mathcal{T}) \sum_{|S|=k,\ S\ni i} \hat{f}(S)^2 \\ &\leq \mathbf{Var}[f] \cdot (\max \delta_i(\mathcal{T})) \cdot \sum_{i=1}^{n} \sum_{|S|=k,\ S\ni i} \hat{f}(S)^2 \\ &= \mathbf{Var}[f] \cdot (\max \delta_i(\mathcal{T})) \cdot k \cdot \sum_{|S|=k} \hat{f}(S)^2,\end{aligned} \tag{9}$$

which is equivalent to (7) after dividing by $\sum_{|S|=k} \hat{f}(S)^2$. Finally, for Corollary 3.4, let $\deg(f)$ denote the "degree" of $f$, $\max\{|S| : f_S \neq 0\}$. Repeat (9) in the more general context $g = \sum_{|S|=k} f_S$, and then sum over $k = 1 \ldots \deg(f)$; one obtains

$$\sum_{k=1}^{\deg(f)} \left(\sum_{|S|=k} \|f_S\|_2^2\right)^2 \leq \mathbf{Var}[f] \cdot \sum_{i=1}^{n} \delta_i(\mathcal{T}) \sum_{S\ni i} \|f_S\|_2^2 = \mathbf{Var}[f] \cdot \sum_{i=1}^{n} \delta_i(\mathcal{T})\ \mathbf{Inf}_i(f).$$

But by Cauchy-Schwarz,

$$\sum_{k=1}^{\deg(f)} \left(\sum_{|S|=k} \|f_S\|_2^2\right)^2 \geq \frac{1}{\deg(f)} \left(\sum_{1\leq|S|\leq k} \|f_S\|_2^2\right)^2 = \frac{1}{\deg(f)} \mathbf{Var}[f]^2.$$

Hence

$$\mathbf{Var}[f] \leq \deg(f) \sum_{i=1}^{n} \delta_i(\mathcal{T})\ \mathbf{Inf}_i(f),$$

which is stronger than Corollary 3.4 since if $f$ has a decision tree in which no path has length more than $k$ then it is easy to see that $\deg(f) \leq k$.

We now present the proof of Theorem A.1:

**Proof:** Without loss of generality we may assume $\mathbf{E}[f] = \mathbf{E}[g] = 0$. We will prove

$$\mathbf{E}[fg] \leq \|f\|_2 \sqrt{\sum_{i=1}^{n} \delta_i(\mathcal{T}) \mathbf{Inf}_i(g)}, \tag{10}$$

and since we can also assume $\mathbf{E}[fg] \geq 0$ by replacing $g$ with $-g$, squaring the above proves the theorem.

The choice of a random $x \leftarrow \Omega$ can be achieved via a two-step procedure: First, pick a random DDT $T$ from $\mathcal{T}$, and pick a random root-to-leaf path $P$ of $T$ according to the natural probability

distribution on $T$'s paths inherited from $\Omega$. This gives a "partial input" that will form part of $x$. Second, choose the remainder of $x$ — call it $x_{\overline{P}}$ — by choosing each of the unset coordinates $x_i$ uniformly from $\mu_i$. Given a function $h : \Omega \to \mathbb{R}$, we denote by $h|_P$ the restricted function given by fixing the partial input from $P$.

We have

$$\mathbf{E}[fg] = \mathop{\mathbf{E}}_{P} \mathop{\mathbf{E}}_{x_{\overline{P}}}[f|_P(x_{\overline{P}})g|_P(x_{\overline{P}})] = \mathop{\mathbf{E}}_{P}[f(P) \mathop{\mathbf{E}}_{x_{\overline{P}}}[g|_P(x_{\overline{P}})]],$$

where we have abused notation by writing $f(P)$ in place of the constant function $f|_P$. By Cauchy-Schwarz we have

$$\mathop{\mathbf{E}}_{P}[f(P) \mathop{\mathbf{E}}_{x_{\overline{P}}}[g|_P(x_{\overline{P}})]] \leq \sqrt{\mathop{\mathbf{E}}_{P}[f(P)^2]} \sqrt{\mathop{\mathbf{E}}_{x_{\overline{P}}}[g|_P(x_{\overline{P}})]^2} = \|f\|_2 \sqrt{\mathop{\mathbf{E}}_{x_{\overline{P}}}[g|_P(x_{\overline{P}})]^2}$$

and so to prove (10) it remains to show

$$\mathop{\mathbf{E}}_{x_{\overline{P}}}[g|_P(x_{\overline{P}})]^2 \leq \sum_{i=1}^{n} \delta_i(\mathcal{T}) \mathbf{Inf}_i(g). \tag{11}$$

For notational ease let us write $G = g|_P$. We will also write the Fourier transform of $G$ as $G = \sum_{S \subseteq [n]} G_S$. This is also a slight abuse of notation, in that $G$ is not a function on all of $[n]$; however, we can let $G_S = 0$ if $S$ includes a coordinate on which $G$ does not depend.

Since $\mathcal{S}$, the antichain on which $g$'s Fourier expansion is supported, cannot contain $\emptyset$, we surely have

$$\mathop{\mathbf{E}}_{x_{\overline{P}}}[g|_P(x_{\overline{P}})]^2 = \|G_\emptyset\|_2^2 \leq \sum_{S \subseteq [n]} \|G_S\|_2^2 - \sum_{S \in \mathcal{S}} \|G_S\|_2^2. \tag{12}$$

By Plancherel and the fact that $g$'s Fourier expansion is supported on $\mathcal{S}$, the first sum on the right in (12) is

$$\mathop{\mathbf{E}}_{x_{\overline{P}}}[G(x_{\overline{P}})^2] = \mathop{\mathbf{E}}_{x \leftarrow \Omega}[g(x)^2] = \sum_{S \in \mathcal{S}} \|g_S\|_2^2.$$

Let's now consider the subtracted sum on the right in (12), $\sum_{S \in \mathcal{S}} \|G_S\|_2^2$. We initially have $g = \sum_{S \in \mathcal{S}} g_S$; then the coordinates in $P$ are fixed and $g$ becomes $G$. Fix a particular $S \in \mathcal{S}$. If $P$ contains a coordinate in $S$ then this "reduces" the component $g_S$ and it will not appear in the Fourier expansion of $G$. On the other hand, if $P$ does not contain any coordinate of $S$, then the component $g_S$ will remain in the Fourier expansion of $G$; here we crucially use the fact that $\mathcal{S}$ is an antichain, so that it is impossible for any component $g_{S'}$ "above" $g_S$ (i.e., with $S' \supsetneq S$) to "reduce" onto $g_S$. Thus we conclude

$$\sum_{S \in \mathcal{S}} \|G_S\|_2^2 = \sum_{S \in \mathcal{S}} \mathbf{Pr}[P \text{ queries no coord. in } S] \cdot \|g_S\|_2^2.$$

Therefore

$$\begin{aligned}
(12) &= \sum_{S \in \mathcal{S}} \|g_S\|_2^2 - \sum_{S \in \mathcal{S}} \mathbf{Pr}[P \text{ queries no coord. in } S] \cdot \|g_S\|_2^2 \\
&= \sum_{S \subseteq [n]} \mathbf{Pr}[P \text{ queries at least one coord. in } S] \cdot \|g_S\|_2^2 \\
&\leq \sum_{S \subseteq [n]} \left( \sum_{i \in S} \delta_i(\mathcal{T}) \right) \cdot \|g_S\|_2^2 \\
&= \sum_{i \in S} \delta_i(\mathcal{T}) \sum_{S \ni i} \|g_S\|_2^2 = \sum_{i \in S} \delta_i(\mathcal{T}) \mathbf{Inf}_i(g),
\end{aligned}$$

confirming (11) and completing the proof. □

Two final remarks: First, it is tempting to wonder if the condition that $g$ is supported on an antichain can be removed. However this is not possible in general, for then we could take $g = f$ and recover Corollary 3.4 without the factor $k$, contradicting the example in Figure 1. Second: it would be interesting to see if Theorem A.1 could be unified with any of the Efron-Stein inequality, Talagrand's inequality, or our main inequality Theorem 3.2. At present we do not see any way do so.

## B Alternate two function version

We now give an alternate two function extension of the theorem.

**Theorem B.1 (Covariance inequality)** *Let $(\Omega, \mu)$ be an $n$-wise product probability space. Let $f : \Omega \to [-1, 1]$, $g : \Omega \to \mathbb{R}$, and let $\mathcal{T}$ be a randomized decision tree computing $f$. Set $\delta_i(f) = \delta_i(\mathcal{T})$ and $\rho_1(x, y) = |x - y|$ for $x, y \in \mathbb{R}$. Then*

$$\big|\mathbf{Cov}[f, g]\big| \leq \sum_{i=1}^{n} \delta_i(f)\, \mathbf{Inf}_i^{\rho_1}[g].$$

**Proof:** Again, we may assume that the tree is deterministic. It suffices to consider the case $\mathbf{Cov}[f, g] \geq 0$, since we may replace $f$ by $-f$. We have $\mathbf{E}[f(x)g(v[0])] = \mathbf{E}[f(v[0])g(v[0])]$, because $f(v[0]) = f(x)$, and also $\mathbf{E}[f(v[0])g(v[0])] = \mathbf{E}[f(x)g(x)]$, because $v[0]$ has the same distribution as $x$. Hence,

$$\begin{aligned}\mathbf{Cov}[f, g] = \mathbf{E}\big[f(x)g(x)\big] - \mathbf{E}\big[f(x)g(y)\big] &= \mathbf{E}\big[f(x)g(v[0]) - f(x)g(v[d])\big] \\ &= \mathbf{E}\Big[f(x)\sum_{t=1}^{d}\big(g(v[t-1]) - g(v[t])\big)\Big] \leq \mathbf{E}\Big[\sum_{t=1}^{d}\big|g(v[t-1]) - g(v[t])\big|\Big].\end{aligned}$$

The rest of the proof proceeds as in Theorem 3.2 and will be omitted. □