Extensions to McDiarmid's inequality when differences are bounded with high probability

Samuel Kutin*

April 12, 2002

Abstract

The method of independent bounded differences (McDiarmid, 1989) gives large-deviation concentration bounds for multivariate functions in terms of the maximum effect that changing one coordinate of the input can have on the output. This method has been widely used in combinatorial applications, and in learning theory. In some recent applications to the theory of algorithmic stability (Kutin and Niyogi, 2002), we need to consider the case where changing one coordinate of the input usually leads to a small change in the output, but not always.

We prove two extensions to McDiarmid's inequality. The first applies when, for most inputs, any small change leads to a small change in the output. The second applies when, for a randomly selected input and a random one-coordinate change, the change in the output is usually small.

1 Introduction

How can we bound the concentration of a random variable about its mean? The classic large-deviation concentration inequality is due to Chernoff [6]:

Theorem 1.1 (Chernoff [6]) *Let $\xi_1, \ldots, \xi_m$ be random variables with $|\xi_k| \leq 1$ and $\mathbf{E}(\xi_k) = 0$ for all k. Let $X = \frac{1}{m}\sum_{k=1}^{m} \xi_k$. Then, for any $\tau > 0$,*

$$\Pr\left(X \geq \tau\right) \leq \exp\left(-\frac{\tau^2 m}{2}\right)$$

There have been a number of generalizations of Chernoff's inequality, such as the Hoeffding-Azuma inequality [9, 1, 7] for martingale difference sequences. In this paper, we discuss one corollary of Hoeffding-Azuma, McDiarmid's method of independent bounded differences [18].

*Department of Computer Science, University of Chicago, 100 E. 58th Street, Chicago, IL 60637. Email: kutin@cs.uchicago.edu.

Definition 1.2 Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω. We say that X is *uniformly difference-bounded by* c if the following holds: for any k, if $\omega, \omega' \in \Omega$ differ only in the kth coordinate, then

$$|X(\omega) - X(\omega')| \leq c. \tag{1}$$

Theorem 1.3 (McDiarmid [18]) *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω which is uniformly difference-bounded by $\frac{\lambda}{m}$. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$,*

$$\Pr(X - \mu \geq \tau) \leq \exp\left(-\frac{2\tau^2 m}{\lambda^2}\right)$$

Our main results are two extensions to McDiarmid's Theorem, which apply when Inequality (1) holds with high probability.

Note 1.4 We assume throughout that the probability measure on Ω is the product of the measures on each Ω_k. So choosing a point in Ω corresponds to choosing each coordinate independently.

We use the notation

$$\forall^{\delta} \omega \in \Omega, \qquad \Phi(\omega)$$

to mean "$\Phi(\omega)$ holds for all but a δ fraction of Ω", or, equivalently, $\Pr_{\omega \in \Omega}(\Phi(\omega)) \geq 1 - \delta$.

We consistently use X, Y, Z for real-valued random variables on Ω (i.e., measurable functions from Ω to $\mathbb{R}$). We use ξ to denote the random variable which corresponds to choosing an element of Ω (so ξ is the identity function on Ω).

We use ω, χ for elements of Ω, and v, x, y for elements of some Ω_k. For the most part, other letters represent real numbers.

Remark 1.5 McDiarmid's Theorem 1.3 implies Chernoff's Theorem 1.1. We take $\Omega_k = [-1, 1]$ and $\lambda = 2$.

McDiarmid credits Maurey's work [14] on functions on permutation spaces as the first example of a result on independent bounded differences. McDiarmid gives a proof [18] of Theorem 1.3 without explicit reference to martingales, and also a proof [19] of Theorem 1.3 and other more general results based on martingale difference sequences. We use some of McDiarmid's general results, and we further generalize another.

McDiarmid [19] catalogs a number of applications of the method of independent bounded differences, including Bollobás's concentration bounds for the chromatic number of a random graph [3]. More recently, Theorem 1.3 has been used by learning theorists: for example, Bousquet and Elisseeff [5] use McDiarmid's theorem to prove that algorithmic stability gives good concentration bounds on generalization error, Freund, et al. [8] use the theorem to prove that the average of a collection of classifiers has good generalization error, and McAllester and Schapire [16, 17] use the theorem to prove concentration bounds for the accuracy of Good-Turing estimators. Theorem 1.3 has also been used [12] to prove concentration bounds for the training error rate of weak learning algorithms.

However, the condition of McDiarmid's Theorem, that Inequality (1) hold for *every* pair of points ω, ω' differing in only one coordinate, is too restrictive for some applications. In particular, the notion of algorithmic stability used by Bousquet and Elisseeff requires that any small change in the training set yields a small change in the final hypothesis of the learning algorithm. Their definition of stability is too rigid to be widely applicable.

We prove two extensions of McDiarmid's Theorem, when Inequality (1) holds most of the time. In both cases, we also require that the function be uniformly distance-bounded by some b, but b can be significantly larger than c. Our extensions allow for a relaxed definition of stability [12, 13], which permits the analysis of a broader collection of learning algorithms within the framework of stability.

Our first extension allows for the possibility of a "bad" set B of inputs for which Inequality (1) does not hold:

Definition 1.6 Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω. We say that X is *strongly difference-bounded by* (b, c, δ) if the following holds: there is a "bad" subset $B \subset \Omega$, where $\delta = \Pr(\omega \in B)$. If $\omega, \omega' \in \Omega$ differ only in the kth coordinate, and $\omega \notin B$, then

$$|X(\omega) - X(\omega')| \leq c.$$

Furthermore, for any ω and ω' differing only in the kth coordinate,

$$|X(\omega) - X(\omega')| \leq b.$$

Our second extension has an even weaker hypothesis:

Definition 1.7 Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω. We say that X is *weakly difference-bounded by* (b, c, δ) if the following holds: for any k,

$$\forall^{\delta}(\omega, \upsilon) \in \Omega \times \Omega_k, \qquad |X(\omega) - X(\omega')| \leq c,$$

where $\omega'_k = \upsilon$ and $\omega'_i = \omega_i$ for $i \neq k$. In words, if we choose $\omega \in \Omega$, and $\upsilon \in \Omega_k$, and we construct ω' by replacing the kth entry of ω with υ, then Inequality (1) holds for all but a δ fraction of the choices. Furthermore, for any ω and ω' differing only in the kth coordinate,

$$|X(\omega) - X(\omega')| \leq b.$$

Note 1.8 The condition of being strongly difference-bounded by (b, c, δ) can be phrased as:

$$\forall^{\delta}\omega \in \Omega, \quad \forall k, \quad \forall \upsilon \in \Omega_k, \qquad |X(\omega) - X(\omega')| \leq c,$$

where ω' is ω with the kth coordinate replaced by υ. Therefore, strong difference-boundedness is strictly stronger than weak difference-boundedness.

Our main result is that either of these notions of difference-boundedness implies a McDiarmid-like concentration bound.

Theorem 1.9 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω which is strongly difference-bounded by $(b, \frac{\lambda}{m}, \exp(-Km))$. Let $\mu = \mathbf{E}(X)$. If $0 < \tau \leq T_1(\lambda, K)$, and $m \geq M_1(b, \lambda, K)$, then,*

$$\Pr(|X - \mu| \geq \tau) \leq 4 \exp\left(-\frac{\tau^2 m}{8\lambda^2}\right).$$

Theorem 1.10 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω which is weakly difference-bounded by $(b, \frac{\lambda}{m}, \exp(-Km))$. Let $\mu = \mathbf{E}(X)$. If $0 < \tau \leq T_2(b, \lambda, K)$, and $m \geq M_2(b, \lambda, K, \tau)$, then*

$$\Pr(|X - \mu| \geq \tau) \leq 4 \exp\left(-\frac{\tau^2 m}{40\lambda^2}\right).$$

Note 1.11 The values of the bounds T_1, M_1, T_2, M_2 for which we prove Theorems 1.9 and 1.10 are as follows:

$$\begin{aligned}
T_1(\lambda, K) &= 2\lambda\sqrt{K} \\
M_1(b, \lambda, K) &= \max\left\{\frac{b}{\lambda}, 3\left(\frac{6}{K} + 3\right)\ln\left(\frac{6}{K} + 3\right)\right\} \\
T_2(b, \lambda, K) &= \min\left\{\frac{15\lambda}{2}, 4\lambda\sqrt{K}, \frac{\lambda^2 K}{b}\right\} \\
M_2(b, \lambda, K, \tau) &= \max\left\{\frac{b}{\lambda}, \lambda\sqrt{40}, 3\left(\frac{24}{K} + 3\right)\ln\left(\frac{24}{K} + 3\right), \frac{1}{\tau}\right\}
\end{aligned}$$

Remark 1.12 The limitation on m in Theorems 1.9 and 1.10 is straightforward: the results hold for sufficiently large m.

In Theorem 1.10, the lower bound on m depends upon τ as well. If we had $m < 1/\tau$, then we would have

$$4 \exp\left(-\frac{\tau^2 m}{40\lambda^2}\right) > 4 \exp\left(-\frac{1}{40 m \lambda^2}\right).$$

For constant λ, this expression approaches 1 as $m \to \infty$. Our interest is in concentration bounds which get tighter, or remain constant, as $m \to \infty$. We would only want to apply the conclusion of Theorem 1.10 when $\tau = \Omega(1/\sqrt{m})$. Hence, requiring $m \geq 1/\tau$ is not a limitation.

Remark 1.13 The upper bound on τ in Theorems 1.9 and 1.10 requires some discussion. In our applications, the uniform difference bound b generally comes from a statement of the form

$$\forall \omega, \chi \in \Omega \qquad |X(\omega) - X(\chi)| \leq b.$$

So, $|X - \mu|$ can never be more than b, and there is no reason to consider $\tau > b$. In practice, the upper bounds $T_1(\lambda, K)$ and $T_2(b, \lambda, K)$ of Note 1.11 are larger than b. So these bounds do not limit the application of our results.

Remark 1.14 In practice, in our applications, the standard deviation of X is roughly $C/\sqrt{m}$. So our results give Chernoff-like bounds when phrased in terms of standard deviation.

Recently, several probabilistic notions of stability have been defined.

Strong hypothesis stability [12, 13] allows for some unlikely training sets to be bad, but requires that, for any good training set, changing any single point leads to a small change in the final hypothesis. Theorem 3.6, a general version of Theorem 1.9, can be used to prove that strong hypothesis stability gives good bounds on generalization error [12, 13]. Theorem 3.6 also implies concentration bounds for the accuracy of Good-Turing estimators [11].

Weak hypothesis stability [13] states that, if we randomly select a training set and then change a single point, this change usually leads to a small change in the final hypothesis. Training stability [13] is an even weaker notion of stability. Theorem 4.8, a general version of Theorem 1.10, can be used to prove that weak hypothesis stability, or even training stability, is also sufficient to give good bounds on generalization error [13].

In Section 2, we define more general notions of difference-boundedness. We also discuss the martingale terminology we need to prove our main results, as well as some lemmas we will use to manipulate the moment generating function e^X. As an example of these lemmas, we prove an extended version of Bernstein's Theorem in Section 2.4.

We prove Theorem 1.9 in Section 3; our proof closely follows a proof of Theorem 3.2 by McDiarmid [19], and uses McDiarmid's Theorem 3.1. We prove Theorem 1.10 in Section 4. In both cases, we state and prove more general versions of the results.

Theorems 1.9 and 1.10 work well in applications to algorithmic stability, but do not apply in some other cases; in particular, when $\lambda = 0$, Theorems 1.9 and 1.10 are vacuous. We give straightforward naïve concentration bounds for the case $\lambda = 0$ in Section 5.

We conclude with some open questions in Section 6.

2 Preliminaries

2.1 Notions of difference-boundedness

In Section 1, we gave several definitions of difference-boundedness. We now extend those definitions, allowing the parameters to vary with k. This will enable us to state and prove our extensions to McDiarmid's inequality in full generality.

Definition 2.1 Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω. We say that X is *uniformly difference-bounded by* $\{c_k\}$ if the following holds: for any k, if $\omega, \omega' \in \Omega$ differ only in the kth coordinate, then

$$|X(\omega) - X(\omega')| \leq c_k.$$

Definition 2.2 Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω. We say that X is *strongly difference-bounded by* $(\{b_k\}, \{c_k\}, \delta)$ if the

following holds: there is a "bad" subset $B \subset \Omega$, where $\delta = \Pr(\omega \in B)$. If $\omega, \omega' \in \Omega$ differ only in the kth coordinate, and $\omega \notin B$, then

$$|X(\omega) - X(\omega')| \leq c_k.$$

Furthermore, for any ω and ω' differing only in the kth coordinate,

$$|X(\omega) - X(\omega')| \leq b_k.$$

Definition 2.3 Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω. We say that X is *weakly difference-bounded by* $(\{b_k\}, \{c_k\}, \{\delta_k\})$ if the following holds: for any k,

$$\forall^{\delta}(\omega, \upsilon) \in \Omega \times \Omega_k, \qquad |X(\omega) - X(\omega')| \leq c_k,$$

where $\omega'_k = \upsilon$ and $\omega'_i = \omega_i$ for $i \neq k$. In words, if we choose $\omega \in \Omega$, and $\upsilon \in \Omega_k$, and we construct ω' by replacing the kth entry of ω with υ, then the inequality holds for all but a δ fraction of the choices. Furthermore, for any ω and ω' differing only in the kth coordinate,

$$|X(\omega) - X(\omega')| \leq b_k.$$

2.2 Martingales

We begin with some general definitions for probability spaces; we follow the notation of McDiarmid [19].

Let Ω be a probability space. A *σ-field* on Ω is a collection $\mathcal{F}$ of subsets of Ω which contains $\emptyset$ and which is closed under complementation and countable union (and, hence, countable intersection). Given any such $\mathcal{F}$, we can partition Ω into disjoint *blocks*, such that $\mathcal{F}$ is the collection of unions of blocks.

An *$\mathcal{F}$-measurable function* on Ω is one which is constant on each block of $\mathcal{F}$. Given a random variable X on Ω, we can construct several natural $\mathcal{F}$-measurable functions:

- $\mathbf{E}(X \mid \mathcal{F})$: the value on each block is the average of X
- $\sup(X \mid \mathcal{F})$: the value on each block is the supremum of X.
- $\operatorname{ran}(X \mid \mathcal{F}) = \sup(X \mid \mathcal{F}) + \sup(-X \mid \mathcal{F})$, the *range* of X.
- $\operatorname{Var}(X \mid \mathcal{F}) = \mathbf{E}(X^2 \mid \mathcal{F}) - \mathbf{E}(X \mid \mathcal{F})^2$.

A *filter* is a nested sequence of σ-fields $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots$. We will be interested in finite filters, with some maximal $\mathcal{F}_m$. A *martingale* is a sequence of random variables $X_0, \ldots, X_m$ for which $X_k = \mathbf{E}(X_{k+1} \mid \mathcal{F}_k)$ for every k. Note that X_k is $\mathcal{F}_k$-measurable. The *martingale difference sequence* $Y_1, \ldots, Y_m$ is given by $Y_k = X_k - X_{k-1}$; then $\mathbf{E}(Y_k \mid \mathcal{F}_{k-1}) = 0$ for all k.

Note 2.4 Given any $\mathcal{F}_m$-measurable random variable X, the sequence $X_k = \mathbf{E}(X \mid \mathcal{F}_k)$ is a martingale; $X_m = X$, and $X_0 = \mathbf{E}(X)$. We have $Y_k = \mathbf{E}(X \mid \mathcal{F}_k) - \mathbf{E}(X \mid \mathcal{F}_{k-1})$.

Note 2.5 Unless otherwise stated, we assume that $\Omega = \prod_{k=1}^{m} \Omega_k$. We let $\mathcal{F}_k$ be the σ-field whose blocks are of the form

$$\{\omega_1\} \times \{\omega_2\} \times \cdots \times \{\omega_k\} \times \Omega_{k+1} \times \cdots \times \Omega_m.$$

A block in $\mathcal{F}_k$ is determined by the first k coordinates, and a function on Ω is $\mathcal{F}_k$-measurable if and only if it depends only on the first k coordinates. The filter corresponds to revealing the coordinates one at a time.

Given a finite filter $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq F_m$, and an $\mathcal{F}_m$-measurable random variable X, define the martingale $\{X_k\}$ and the martingale difference sequence $\{Y_k\}$ as in Note 2.4. We now define the following $\mathcal{F}_{k-1}$-measurable functions:

$$\begin{aligned} \mathrm{ran}_k &= \mathrm{ran}(Y_k \mid \mathcal{F}_{k-1}) = \mathrm{ran}(X_k \mid \mathcal{F}_{k-1}) \\ \mathrm{Var}_k &= \mathrm{Var}(Y_k \mid \mathcal{F}_{k-1}) = \mathrm{Var}(X_k \mid \mathcal{F}_{k-1}) \\ \mathrm{dev}_k^+ &= \sup(Y_k \mid \mathcal{F}_{k-1}) \\ p_{k,d} &= \Pr(Y_k > d \mid \mathcal{F}_{k-1}) \end{aligned}$$

Note that $p_{k,d}$ depends on an additional parameter d.

We now let $\max \mathrm{dev}^+$ denote the maximum of $\sup(\mathrm{dev}_k^+)$ over all k. We also define the following random variables:

- The *sum of squared conditional ranges* $R^2 = \sum_{k=1}^{m} \mathrm{ran}_k^2$.
- The *sum of conditional variances* $V = \sum_{k=1}^{m} \mathrm{Var}_k$.
- $P_d = \sum_{k=1}^{m} p_{k,d}$.

Note 2.6 Suppose we are in the setting of Note 2.5: Our probability space is a product $\Omega = \prod_{k=1}^{m} \Omega_k$, and we are given a random variable X on Ω. We define $\{X_k\}$ to be the martingale which corresponds to revealing one coordinate at a time. The definitions above can now be phrased as follows: fix some k, and any $\omega_1, \ldots, \omega_{k-1}$ with $\omega_i \in \Omega_i$. Let ξ_i be the random variable corresponding to choosing from Ω_i, and write $\xi = (\xi_1, \ldots, \xi_m)$. Let Γ denote the event that $\xi_i = \omega_i$ for $1 \leq i \leq k-1$. We define $\phi \colon \Omega_k \to \mathbb{R}$ by

$$\phi(x) = \mathbf{E}(X(\xi) \mid \Gamma, \xi_k = x) - \mathbf{E}(X(\xi) \mid \Gamma).$$

We can now define $\mathrm{ran}_k = \mathrm{ran}(\phi)$, $\mathrm{Var}_k = \mathrm{Var}(\phi)$, $\mathrm{dev}_k^+ = \sup(\phi)$, and $p_{k,d} = \Pr(\phi > d)$. We define $\max \mathrm{dev}^+$, R^2, V, and P_d as above. In this case, we call R^2 the *sum of squared ranges*, and V the *sum of variances*.

2.3 The moment generating function

We now state some lemmas we will need to manipulate the *moment generating function* e^X of a random variable X.

The first lemma is due to Steiger [20], though our notation follows that of McDiarmid [19, Lemma 2.8].

Lemma 2.7 (Steiger [20]) *Let*

$$\begin{aligned} g(z) &= \frac{1}{2} + \frac{z}{3!} + \frac{z^2}{4!} + \cdots \\ &= \frac{e^z - 1 - z}{z^2} \quad \text{if } z \neq 0. \end{aligned}$$

Then the function g is increasing. If X is a random variable satisfying $\mathbf{E}(X) = 0$ *and* $X \leq d$, *then*

$$\mathbf{E}(e^X) \leq \exp(g(d)\,\mathrm{Var}(X)).$$

We will need a slight generalization of this result:

Lemma 2.8 *Let* $g(z)$ *be the function of Lemma 2.7.*

1. *The function g is increasing.*
2. *Let X be a random variable satisfying* $\mathbf{E}(X) = 0$ *and* $X \leq D$. *For any* $d < D$, *if* $\delta = \Pr(X > d)$, *then*
$$\mathbf{E}(e^X) \leq \exp(g(d)\,\mathrm{Var}(X) + \delta e^D).$$

Proof: For completeness, we begin with a proof of Part 1.

Proof of Part 1: For $z \neq 0$, we have

$$g'(z) = \frac{(z-2)e^z + z + 2}{z^3}.$$

To prove $g(z)$ is increasing, it suffices to prove that $g'(z) > 0$ for all $z \neq 0$. Let $f(z) = z^3 g'(z) = (z-2)e^z + z + 2$; we will show that $f(z) > 0$ when $z > 0$ and $f(z) < 0$ when $z < 0$.

First, we note that $f(0) = 0$. Next, we observe that, for any z,

$$f'(z) = (z-1)e^z + 1 \geq (z-1)(z+1) + 1 = z^2 \geq 0,$$

with equality only if $z = 0$. We conclude that $f(z)$ is increasing, so $f(z) > 0$ for $z > 0$ and $f(z) < 0$ for $z < 0$. By the above reasoning, $g(z)$ is increasing. □

Proof of Part 2: For any z, $e^z = 1 + z + z^2 g(z)$. So,

$$\begin{aligned} \mathbf{E}(e^X) &= 1 + \mathbf{E}(X) + \mathbf{E}(X^2 g(d)) + \mathbf{E}(X^2(g(X) - g(d))) \\ &= 1 + g(d)\,\mathrm{Var}(X) + \mathbf{E}(X^2(g(X) - g(d))). \end{aligned}$$

Let $Y = X^2(g(X) - g(d))$. By Lemma 2.7, $g(X)$ is increasing, so, when $X \leq d$, $Y \leq 0$. Also, since $g(d) > 0$, we always have

$$Y \leq D^2(g(D) - g(d)) \leq D^2 g(D) \leq e^D.$$

So, $\mathbf{E}(Y) \leq \delta e^D$. Hence,

$$\mathbf{E}(e^X) \leq 1 + g(d)\,\mathrm{Var}(X) + \delta e^D \leq \exp(g(d)\,\mathrm{Var}(X) + \delta e^D).$$

□

This concludes the lemma. ■

As an example of our technique, we will use Lemma 2.8 to prove an extended version of Bernstein's Theorem in Section 2.4.

The next lemma, due to McDiarmid [19], is based on Lemma 3.4 of Kahn [10].

Lemma 2.9 (McDiarmid [19, Lemma 3.16]) *Let $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_m$ be a filter, and let $Y_1, \ldots, Y_m$ be a corresponding martingale difference sequence, where each Y_k is bounded above. Let the random variable Z be the indicator of some event. Then, for any h,*

$$\mathbf{E}\left(Ze^{h\sum_k Y_k} \mid \mathcal{F}_0\right) \leq \sup\left(\prod_k \mathbf{E}(e^{hY_k} \mid \mathcal{F}_{k-1}) \mid \mathcal{F}_0\right).$$

Finally, we use a technical lemma of McDiarmid [19]:

Lemma 2.10 (McDiarmid [19, Lemma 2.4]) *For all $z \geq 0$,*

$$(1+z)\ln(1+z) - z \geq \frac{3z^2}{6+2z}.$$

2.4 Bernstein's Theorem

Bernstein's Theorem is a Chernoff-like large-deviation concentration inequality. McDiarmid [19] gives a proof of Bernstein's Theorem using Lemma 2.7.

In this section, we state and prove Theorem 2.12, an extension to Bernstein's Theorem. We use the same approach McDiarmid uses to prove Bernstein's Theorem, but we use Lemma 2.8 in place of Lemma 2.7.

The proof of Theorem 2.12 illustrates the power of Lemma 2.8. It also illustrates the primary technique of Section 4: our proof of Theorem 1.10 follows the same lines as McDiarmid's proof of Theorem 1.3, except that we use Lemma 2.8 in place of Lemma 2.7.

We begin with a statement of Bernstein's Theorem (see, for example, Bennett [2]):

Theorem 2.11 (Bernstein) *Let $\xi_1, \ldots, \xi_m$ be independent random variables, with $\xi_k - \mathbf{E}(\xi_k) \leq d$ for all k. Let $X = \sum_{k=1}^m \xi_k$, and let $V = \mathrm{Var}(X) = \sum_{k=1}^m \mathrm{Var}(\xi_k)$.*

Let $\mu = \mathbf{E}(X)$. For any $\tau \geq 0$,

$$\Pr(X - \mu \geq \tau) \leq \exp\left(-\frac{V}{d^2}((1+\epsilon)\ln(1+\epsilon) - \epsilon)\right) \tag{2}$$

$$\leq \exp\left(\frac{-\tau^2}{2V\left(1+\frac{\epsilon}{3}\right)}\right), \tag{3}$$

where $\epsilon = d\tau/V$.

In many applications, the error term ϵ in Inequalities (2) and (3) is negligible. For example, if each ξ_i is chosen uniformly from $\{-1, 1\}$, then $d = 1$ and $V = m$, so for $\tau = o(m)$ we get

$$\Pr(X - \mu \geq \tau) \leq \exp\left(\frac{-\tau^2}{2m}(1 + o(1))\right),$$

which is log-asymptotic to the bound obtained from Chernoff's Theorem 1.1.

However, suppose we are instead given that $\xi_k - \mathbf{E}(\xi_k) \leq d$ with high probability, and $\xi_k - \mathbf{E}(\xi_k) \leq D$ always. A bound of the form

$$\Pr(X - \mu \geq \tau) \leq \exp\left(\frac{-\tau^2}{2V}(1 + o(1))\right)$$

would still be useful (for example, if $|\xi_k| \leq 1$ with high probability, V will still be roughly m). However, the error term $D\tau/V$ may no longer be negligible. We would like to prove a version of Theorem 2.11 which applies in this context and where we still have $\epsilon = d\tau/V$.

McDiarmid [19] gives a proof of Bernstein's Theorem using Lemmas 2.7 and 2.10. We use this same argument, but we use Lemma 2.8 in place of Lemma 2.7.

Theorem 2.12 *Let $\xi_1, \ldots, \xi_m$ be independent random variables. Suppose that, for all k, $\Pr(\xi_k - \mathbf{E}(\xi_k) > d) \leq \delta$ and $\xi_k - \mathbf{E}(\xi_k) \leq D$. Let $X = \sum_{k=1}^m \xi_k$, and let $V = \mathrm{Var}(X) = \sum_{k=1}^m \mathrm{Var}(\xi_k)$.*

Let $\mu = \mathbf{E}(X)$. For any $\tau \geq 0$,

$$\Pr(X - \mu \geq \tau) \leq \exp\left(-\frac{V}{d^2}((1+\epsilon)\ln(1+\epsilon) - \epsilon) + m\delta(1+\epsilon)^{D/d}\right) \tag{4}$$

$$\leq \exp\left(\frac{-\tau^2}{2V\left(1 + \frac{\epsilon}{3}\right)} + m\delta(1+\epsilon)^{D/d}\right), \tag{5}$$

where $\epsilon = d\tau/V$.

Proof: For any $h > 0$, by Lemma 2.8,

$$\begin{aligned}
\mathbf{E}\left(e^{h(X-\mu)}\right) &= \prod_{k=1}^m \mathbf{E}\left(e^{h(\xi_k - \mathbf{E}(\xi_k))}\right) \\
&\leq \prod_{k=1}^m \exp\left(g(hd)\,\mathrm{Var}(h\xi_k) + \delta e^{hD}\right) \\
&= \exp\left(g(hd)h^2V + m\delta e^{hD}\right).
\end{aligned}$$

So, for any $\tau \geq 0$, by Markov's inequality,

$$\begin{aligned}
\Pr(X - \mu \geq \tau) &= \Pr\left(e^{h(X-\mu)} \geq e^{h\tau}\right) \\
&\leq e^{-h\tau}\mathbf{E}\left(e^{h(X-\mu)}\right) \\
&\leq \exp\left(-h\tau + g(hd)h^2V + m\delta e^{hD}\right).
\end{aligned}$$

The expression $-h\tau + g(hd)h^2V$ is minimized at $h = \frac{1}{d}\ln(1+\epsilon)$, where $\epsilon = d\tau/V$. This value of h gives us Inequality (4).

Inequality (5) now follows from Lemma 2.10. ■

3 Strongly difference-bounded functions

In this section, we prove Theorem 1.9. We follow the same general argument McDiarmid uses to prove Theorem 1.3. Both his proof and ours use the following result about sums of squared ranges [19, Theorem 3.7]:

Theorem 3.1 (McDiarmid) *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω. Let $\hat{r}^2$ denote the maximum sum of squared ranges $\sup(R^2)$. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau \geq 0$,*

$$\Pr(X - \mu \geq \tau) \leq \exp(-2\tau^2/\hat{r}^2).$$

More generally, let B be a "bad" subset of Ω such that $R^2(\omega) \leq r^2$ for each $\omega \notin B$. Then

$$\Pr(X - \mu \geq \tau) \leq \exp(-2\tau^2/r^2) + \Pr(\omega \in B).$$

McDiarmid [19] uses Theorem 3.1 to prove the following theorem, of which Theorem 1.3 is a special case:

Theorem 3.2 (McDiarmid [18]) *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is uniformly difference-bounded by $\{c_k\}$. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$,*

$$\Pr(X - \mu \geq \tau) \leq \exp\left(\frac{-2\tau^2}{\sum c_k^2}\right)$$

We now prove our most general form of Theorem 1.9.

Theorem 3.3 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is strongly difference-bounded by $(\{b_k\}, \{c_k\}, \delta)$. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$, and any $\alpha_1, \ldots, \alpha_m > 0$,*

$$\Pr(|X - \mu| \geq \tau) \leq 2\left(\exp\left(\frac{-\tau^2}{2\sum_k (c_k + b_k\alpha_k)^2}\right) + \delta \sum_k \frac{1}{\alpha_k}\right). \tag{6}$$

Proof: We let ξ_i be the random variable corresponding to choosing from Ω_i, and we write $\xi = (\xi_1, \ldots, \xi_m)$. Fix some k and $\omega_- = (\omega_1, \ldots, \omega_{k-1}) \in \prod_{i=1}^{k-1} \Omega_i$, and let Γ be the event that $\xi_i = \omega_i$ for $1 \leq i \leq k-1$.

We are interested in bounding $\mathrm{ran}(\omega_1, \ldots, \omega_{k-1})$. Choose some $x \in \Omega_k$. For any $\omega_+ \in \prod_{i=k+1}^m \Omega_i$, if $(\omega_-, x, \omega_+) \notin B$, then, for every $y \in \Omega_k$,

$$|X(\omega_-, x, \omega_+) - X(\omega_-, y, \omega_+)| \leq c_k.$$

Otherwise, if $(\omega_-, x, \omega_+) \in B$, we still have

$$|X(\omega_-, x, \omega_+) - X(\omega_-, y, \omega_+)| \leq b_k$$

for every $y \in \Omega_k$.

So, if $p = \Pr(\xi \in B \mid \Gamma, \xi_k = x)$, then, for any $y \in \Omega_k$,

$$\mathbf{E}_{\omega_+}(|X(\omega_-, x, \omega_+) - X(\omega_-, y, \omega_+)|) \leq (1-p)c_k + pb_k \leq c_k + pb_k.$$

Hence, for any $y, y' \in \Omega_k$,

$$\mathbf{E}_{\omega_+}(|X(\omega_-, y, \omega_+) - X(\omega_-, y', \omega_+|) \leq 2(c_k + pb_k),$$

which implies that

$$\operatorname{ran}(\omega_1, \ldots, \omega_{k-1}) \leq 2(c_k + pb_k).$$

By total probability, there is some way to choose x such that

$$p = \Pr(\xi \in B \mid \Gamma, \xi_k = x) \leq \Pr(\xi \in B \mid \Gamma),$$

and therefore

$$\operatorname{ran}(\omega_1, \ldots, \omega_{k-1}) \leq 2(c_k + b_k \Pr(\xi \in B \mid \Gamma)).$$

Now, we wish to bound the probability that $\Pr(\xi \in B \mid \Gamma)$ is large. Let C_k be the subset of $\prod_{i=1}^{k-1} \Omega_i$ consisting of "bad starts:"

$$C_k = \{(\omega_1, \ldots, \omega_{k-1}) | \Pr(\xi \in B \mid \Gamma) > \alpha_k\}.$$

We let $B_k = C_k \times \Omega_k \times \cdots \times \Omega_m$ be the set of all points which have a bad start; note that $B_k \subset \Omega$. We have $\omega = (\omega_1, \ldots, \omega_m) \in B_k$ if and only if $(\omega_1, \ldots, \omega_{k-1}) \in C_k$.

We observe that

$$\begin{aligned}
\delta = \Pr(\xi \in B) &= \Pr(\xi \in B_k) \Pr(\xi \in B \mid \xi \in B_k) \\
&\geq \Pr(\xi \in B_k) \inf_{(\omega_1, \ldots, \omega_{k-1}) \in C_k} \Pr(\xi \in B \mid \Gamma) \\
&\geq \Pr(\xi \in B_k) \alpha_k,
\end{aligned}$$

and hence

$$\Pr(\xi \in B_k) \leq \delta / \alpha_k.$$

If we define $B' = \bigcup_k B_k$, then

$$\Pr(\xi \in B') \leq \delta \sum_{k=1}^{m} \frac{1}{\alpha_k}.$$

Now, if $\omega \notin B'$, we see that

$$\begin{aligned}
R^2(\omega) &= \sum_{k=1}^{m} (\operatorname{ran}(\omega_1, \ldots, \omega_{k-1}))^2 \\
&\leq \sum_{k=1}^{m} (2(c_k + b_k \Pr(X \in B \mid \Gamma)))^2 \\
&\leq 4 \sum_{k=1}^{m} (c_k + b_k \alpha_k)^2.
\end{aligned}$$

The result now follows immediately from Theorem 3.1. ■

Corollary 3.4 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is strongly difference-bounded by (b, c, δ). Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$, and any $\alpha > 0$,*

$$\Pr(|X - \mu| \geq \tau) \leq 2\left(\exp\left(\frac{-\tau^2}{2m(c + b\alpha)^2}\right) + \frac{m}{\alpha}\delta\right). \tag{7}$$

Proof: This is simply Theorem 3.3 with $b_k = b$, $c_k = c$, and $\alpha_k = \alpha$. ■

The next question is how best to choose the parameter α_k in Inequality (6), or the parameter α in Inequality (7). In applications to algorithmic stability [12, 13], $b_k = \Theta(1)$, $c_k = \Theta(1/m)$, and $\delta = \exp(-\Omega(m))$, so choosing $\alpha_k = c_k/b_k$ is close to optimal:

Corollary 3.5 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is strongly difference-bounded by $(\{b_k\}, \{c_k\}, \delta)$. Assume $b_k \geq c_k > 0$ for all k. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$,*

$$\Pr(|X - \mu| \geq \tau) \leq 2\left(\exp\left(\frac{-\tau^2}{8\sum_k c_k^2}\right) + \delta\sum_k \frac{b_k}{c_k}\right).$$

Proof: Set $\alpha_k = c_k/b_k$ and apply Theorem 3.3. ■

Theorem 3.6 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is strongly difference-bounded by (b, c, δ). Assume $b \geq c > 0$. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$,*

$$\Pr(|X - \mu| \geq \tau) \leq 2\left(\exp\left(\frac{-\tau^2}{8mc^2}\right) + \frac{mb\delta}{c}\right).$$

Proof: Theorem 3.6 is simply Corollary 3.4 with $\alpha = c/b$. ■

We are almost ready to prove Theorem 1.9; we first prove the following technical lemma.

Lemma 3.7 *For any $z > 0$, if $m \geq 3(z + 3)\ln(z + 3)$, then $\frac{m}{\ln m} > z$.*

Proof: We first note that

$$\frac{d}{dm}\frac{m}{\ln m} = \frac{\ln m - 1}{\ln^2 m},$$

so $\frac{m}{\ln m}$ is increasing when $m > e$.

Next, we know $\ln(z + 3) \geq \ln\ln(z + 3)$. Also, $z > 0$, so $\ln(z + 3) > \ln 3$. Hence,

$$\begin{aligned}\frac{m}{\ln m} &\geq \frac{3(z + 3)\ln(z + 3)}{\ln 3 + \ln(z + 3) + \ln\ln(z + 3)} \\ &> \frac{3(z + 3)\ln(z + 3)}{3\ln(z + 3)} = z + 3 > z.\end{aligned}$$

■

Proof of Theorem 1.9: We use Theorem 3.6 with $c = \lambda/m$ and $\delta = \exp(-Km)$. For any $\tau \geq 0$, we have

$$\Pr(|X - \mu| \geq \tau) \leq 2\left(\exp\left(-\frac{\tau^2 m}{8\lambda^2}\right) + m^2\frac{b}{\lambda}\exp(-Km)\right).$$

By our choice of m (see Note 1.11), and by Lemma 3.7, $\frac{m}{\ln m} > \frac{6}{K}$, so $3\ln m < \frac{K}{2}m$. Also, $m \geq \frac{b}{\lambda}$. Therefore,

$$m^2\frac{b}{\lambda}\exp(-Km) \leq m^3\exp(-Km) = \exp(-Km + 3\ln m) \leq \exp\left(-\frac{K}{2}m\right),$$

which implies

$$\Pr(|X - \mu| \geq \tau) \leq 2\left(\exp\left(-\frac{\tau^2 m}{8\lambda^2}\right) + \exp\left(-\frac{K}{2}m\right)\right).$$

Finally, when $\tau \leq 2\lambda\sqrt{K}$, we have

$$-\frac{\tau^2 m}{8\lambda^2} \geq -\frac{K}{2}m,$$

and thus

$$\Pr(|X - \mu| \geq \tau) \leq 4\exp\left(-\frac{\tau^2 m}{8\lambda^2}\right).$$

■

4 Weakly difference-bounded functions

The proof of Theorem 1.10 is more involved than that of Theorem 1.9. We first state, and generalize, a theorem of McDiarmid about martingales [19, Theorem 3.15].

Theorem 4.1 (McDiarmid [19]) *Let X be a random variable with $\mathbf{E}(X) = \mu$, and let $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$ be a filter in $\mathcal{F}$. Let $d = \max \mathrm{dev}^+$, the maximum conditional positive deviation (and assume that d is finite). Then, for any $\tau \geq 0$,*

$$\Pr(X - \mu \geq \tau) \leq \exp\left(\frac{-\tau^2}{2\hat{v}\left(1 + \frac{d\tau}{3\hat{v}}\right)}\right),$$

where $\hat{v}$ is the maximum sum of conditional variances (which is also assumed to be finite). More generally, for any $\tau \geq 0$ and any $v \geq 0$,

$$\Pr(X - \mu \geq \tau) \leq \exp\left(\frac{-\tau^2}{2v\left(1 + \frac{d\tau}{3v}\right)}\right) + \Pr(V(X) > v).$$

Corollary 4.2 (McDiarmid [19, Theorem 3.8]) *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω. Let $d = \max \operatorname{dev}^+$, and let $\hat{v}$ denote the maximum sum of variances $\sup(V)$. (Assume that both d and $\hat{v}$ are finite.) Let $\mu = \mathbf{E}(X)$. Then, for any $\tau \geq 0$,*

$$\Pr(X - \mu \geq \tau) \leq \exp\left(\frac{-\tau^2}{2\hat{v}\left(1 + \frac{d\tau}{3\hat{v}}\right)}\right).$$

More generally, for any $\tau \geq 0$ and any $v \geq 0$,

$$\Pr(X - \mu \geq \tau) \leq \exp\left(\frac{-2\tau^2}{2v\left(1 + \frac{d\tau}{3v}\right)}\right) + \Pr(V > v).$$

Theorem 4.1 and Corollary 4.2 are generalizations of Bernstein's Theorem (see Theorem 2.11).

Theorem 4.1 is not quite sufficient for our needs. We need to consider the case where each martingale difference Y_k is usually bounded by d, but where $\max \operatorname{dev}^+$ is actually a larger value D. We prove the following generalization:

Theorem 4.3 *Let X be a random variable with $\mathbf{E}(X) = \mu$, and let $(\emptyset, \Omega) = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n$ be a filter in $\mathcal{F}$. Let $D = \max \operatorname{dev}^+$, the maximum conditional positive deviation (and assume that D is finite). Then, for any $\tau, d, v, p > 0$,*

$$\Pr(X - \mu \geq \tau) \leq \exp\left(\frac{-\tau^2}{2v\left(1 + \frac{d\tau}{3v}\right)} + p\left(1 + \frac{d\tau}{v}\right)^{D/d}\right) + \Pr((V > v) \vee (P_d > p)).$$

Theorem 4.3 is an extension of McDiarmid's Theorem 4.1 in the same way that Theorem 2.12 is an extension of Bernstein's Theorem 2.11. As in Section 2.4, our proof follows the same lines as McDiarmid's, except that we use Lemma 2.8 in place of Lemma 2.7.

Proof: Let $Y_1, \ldots, Y_m$ be the corresponding martingale difference sequence. Let B denote the bad set where either $V > v$ or $P_d > p$, and let Z be the indicator variable for the event $\Omega \setminus B$. Note that $0 \leq ZV \leq v$ and $0 \leq ZP_d \leq p$. Let $g(z)$ be the function of Lemma 2.7.

By Lemma 2.8, for any $h > 0$,

$$\mathbf{E}(e^{hY_k} \mid \mathcal{F}_{k-1}) \leq \exp(h^2 g(hd) \operatorname{Var}_k + p_{k,d} e^{hD}).$$

So, by Lemma 2.9,

$$\begin{aligned}
\mathbf{E}(Ze^{h(X-\mu)}) &\leq \sup\left(Z \prod_{k=1}^{m} \exp(h^2 g(hd) \operatorname{Var}_k + p_{k,d} e^{hD})\right) \\
&= \sup(Z \exp(h^2 g(hd) V + P_d e^{hD})) \\
&\leq \exp(h^2 g(hd) \sup(ZV) + \sup(ZP_d) e^{hD}) \\
&\leq \exp(h^2 g(hd) v + p e^{hD}).
\end{aligned}$$

Hence, by Markov's inequality,

$$\begin{aligned}\Pr((X-\mu\geq\tau)\wedge(Z=1)) &= \Pr(Ze^{h(X-\mu)}\geq e^{h\tau})\\ &\leq e^{-h\tau}\mathbf{E}(Ze^{h(X-\mu)})\\ &\leq \exp(-h\tau+h^2g(hd)v+pe^{hD}).\end{aligned}$$

Let $\epsilon = d\tau/v$. The expression $-h\tau+h^2g(hd)v$ is minimized when $h=\frac{1}{d}\ln(1+\epsilon)$, yielding

$$\Pr((X-\mu\geq\tau)\wedge(Z=1)) \leq \exp\left(\frac{-v}{d^2}((1+\epsilon)\ln(1+\epsilon)-\epsilon)+pe^{\frac{D}{d}\log(1+\epsilon)}\right).$$

The theorem now follows from Lemma 2.10. ■

Corollary 4.4 *Let $\Omega_1,\ldots,\Omega_m$ be probability spaces. Let $\Omega=\prod_{k=1}^m\Omega_k$, and let X be a random variable on Ω. Let $D=\max\mathrm{dev}^+$, and assume that D is finite. Let $\mu=\mathbf{E}(X)$. Then, for any $\tau,d,v,p>0$,*

$$\Pr(X-\mu\geq\tau)\leq\exp\left(\frac{-\tau^2}{2v\left(1+\frac{d\tau}{3v}\right)}+p\left(1+\frac{d\tau}{v}\right)^{D/d}\right)+\Pr((V>v)\vee(P_d>p)).$$

We are now ready to prove the general version of Theorem 1.10:

Theorem 4.5 *Let $\Omega_1,\ldots,\Omega_m$ be probability spaces. Let $\Omega=\prod_{k=1}^m\Omega_k$, and let X be a random variable on Ω which is weakly difference-bounded by $(\{b_k\},\{c_k\},\{\delta_k\})$. Let $\mu=\mathbf{E}(X)$. Then, for any $\tau>0$, and any positive numbers γ_k, θ_k,*

$$\Pr(X-\mu\geq\tau)\leq\exp\left(\frac{-\tau^2}{2v\left(1+\frac{d\tau}{3v}\right)}+\left(1+\frac{d\tau}{v}\right)^{D/d}\sum_{k=1}^m\frac{\gamma_k}{\theta_k}\right)+\sum_k\frac{\delta_k}{\gamma_k} \tag{8}$$

where

$$\begin{aligned}v &= \sum_k\left((c_k+\theta_kb_k)^2+\frac{\gamma_kb_k^2}{\theta_k}\right),\\ d &= \max_{1\leq k\leq m}\{c_k+\theta_kb_k\},\\ D &= \max_{1\leq k\leq m}\{b_k\}.\end{aligned}$$

Proof: Let $d=\max_k\{c_k+\theta_kb_k\}$, and let $D=\max_k\{b_k\}$. We note that, for any $\omega_1,\ldots,\omega_{k-1}$, we have $\mathrm{dev}_k^+(\omega_1,\ldots,\omega_{k-1})\leq b_k$. So

$$\max\mathrm{dev}^+\leq\max_{1\leq k\leq m}\{b_k\}=D.$$

Fix some k. We say that $\omega_-=(\omega_1,\ldots,\omega_{k-1})$ is "good" if

$$\forall^{\gamma_k}(x,y,\omega_+)\in\Omega_k\times\Omega_k\times\prod_{i=k+1}^m\Omega_i,\qquad |X(\omega_-,x,\omega_+)-X(\omega_-,y,\omega_+)|\leq c_k.$$

Let ζ be the probability that ω_- is bad. Then, with probability at least $\zeta\gamma_k$, we have $|X(\omega_-, x, \omega_+) - X(\omega_-, y, \omega_+)| > c_k$. We conclude that $\zeta\gamma_k \leq \delta_k$, or $\zeta \leq \delta_k/\gamma_k$.

Now, assume ω_- is good. We say that $x \in X_k$ is "good" if

$$\forall^{\theta_k}(y, \omega_+) \in \Omega_k \times \prod_{i=k+1}^{m} \Omega_i, \qquad |X(\omega_-, x, \omega_+) - X(\omega_-, y, \omega_+)| \leq c_k.$$

Let ι be the probability that x is bad. Then

$$\iota\theta_k \leq \Pr_{x,y,\omega_+}(|X(\omega_-, x, \omega_+) - X(\omega_-, y, \omega_+)| \leq c_k) \leq \gamma_k,$$

so $\iota \leq \gamma_k/\theta_k$.

Let $\phi(x)$ be the function of Note 2.6. If x is good, then

$$\begin{aligned} |\phi(x)| &= |\mathbf{E}_{\omega_+}(X(\omega_-, x, \omega_+)) - \mathbf{E}_{y,\omega_+}(X(\omega_-, y, \omega_+))| \\ &\leq \mathbf{E}_{y,\omega_+}(|X(\omega_-, x, \omega_+) - X(\omega_-, y, \omega_+)|) \\ &\leq c_k + \theta_k b_k. \end{aligned}$$

For any x, $|\phi(x)| \leq b_k$. So, for any good ω_-,

$$\mathrm{Var}_k(\omega_-) = \mathbf{E}(\phi(x)^2) \leq (c_k + \theta_k b_k)^2 + \frac{\gamma_k b_k^2}{\theta_k}.$$

Also, since $c_k + \theta_k b_k \leq d$, we have $\phi(x) \leq d$ whenever x is good, and hence

$$p_{k,d}(\omega_-) \leq \Pr(x \text{ is bad}) \leq \frac{\gamma_k}{\theta_k}.$$

So, let

$$\begin{aligned} v &= \sum_{k=1}^{m} \left((c_k + \theta_k b_k)^2 + \frac{\gamma_k b_k^2}{\theta_k} \right). \\ p &= \sum_{k=1}^{m} \frac{\gamma_k}{\theta_k}. \end{aligned}$$

If, for each k, $(\omega_1, \ldots, \omega_{k-1})$ is good, then $V(\omega) \leq v$ and $P_d(\omega) \leq p$. The probability that ω is bad for some k is at most $\sum_k \frac{\delta_k}{\gamma_k}$. The result now follows immediately from Theorem 4.3. ■

Corollary 4.6 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is weakly difference-bounded by (b, c, δ). Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$, and any positive numbers γ, θ,*

$$\Pr(|X - \mu| \geq \tau) \leq 2\left(\exp\left(\frac{-\tau^2}{2v\left(1 + \frac{\tau(c+\theta b)}{3v}\right)} + \frac{m\gamma}{\theta}\left(1 + \frac{\tau(c+\theta b)}{v}\right)^{\frac{b}{c+\theta b}} \right) + \frac{m\delta}{\gamma} \right), \quad (9)$$

where

$$v = m\left((c + \theta b)^2 + \frac{\gamma b^2}{\theta} \right).$$

Proof: This is simply Theorem 4.5 with $b_k = b$, $c_k = c$, $\delta_k = \delta$, $\gamma_k = \gamma$, and $\theta_k = \theta$. This implies that $\max_k\{b_k\} = b$ and $\max_k\{c_k + \theta_k b_k\} = c + \theta b$. ■

We next consider how to choose the parameters γ_k and θ_k in Inequality (8), or γ and θ in Inequality (9). As in Section 3, in applications to algorithmic stability [13], $b_k = \Theta(1)$, $c_k = \Theta(1/m)$, and $\delta_k = \exp(-\Omega(m))$.

Corollary 4.7 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is weakly difference-bounded by $(\{b_k\}, \{c_k\}, \{\delta_k\})$. Assume $b_k \geq c_k > 0$ for all k, and assume $\delta_k \leq (c_k/b_k)^6$ for all k. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$,*

$$\Pr(|X - \mu| \geq \tau) \leq 2\left(\exp\left(\frac{-\tau^2}{10\sum_k c_k^2\left(1 + \frac{d\tau}{15\sum_k c_k^2}\right)} + \left(1 + \frac{d\tau}{4\sum_k c_k^2}\right)^{D/d}\sum_{k=1}^m \frac{b_k\delta_k^{1/2}}{c_k}\right) + \sum_k \delta_k^{1/2}\right),$$

where $d = 2\max_k\{c_k\}$ and $D = \max_k\{b_k\}$.

Proof: Set $\theta_k = c_k/b_k$ and $\gamma_k = \delta_k^{1/2}$ and apply Theorem 4.5. By our assumption on δ_k, we know $\gamma_k \leq (c_k/b_k)^3$, so $4\sum_k c_k^2 \leq v \leq 5\sum_k c_k^2$. ■

Theorem 4.8 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω which is weakly difference-bounded by (b, c, δ). Assume $b \geq c > 0$, and assume $\delta \leq (c/b)^6$. Let $\mu = \mathbf{E}(X)$. Then, for any $\tau > 0$,*

$$\Pr(|X - \mu| \geq \tau) \leq 2\left(\exp\left(\frac{-\tau^2}{10mc^2\left(1 + \frac{2\tau}{15mc}\right)} + \frac{mb\delta^{1/2}}{c}\exp\left(\frac{\tau b}{4mc^2}\right)\right) + m\delta^{1/2}\right).$$

Proof: To prove Theorem 4.8, we use Corollary 4.7 with $b_k = b$ and $c_k = c$, and hence $\theta_k = c/b$ and $\gamma_k = \delta^{1/2}$. Note that $d = 2c$ and $D = b$.

We also use the fact that $1 + (\tau/2mc) \leq \exp(\tau/2mc)$. ■

Proof of Theorem 1.10: We use Theorem 4.8 with $c = \lambda/m$ and $\delta = \exp(-Km)$. Note that, by our choice of m (see Note 1.11), and by Lemma 3.7, $\frac{m}{\ln m} > \frac{24}{K} > \frac{12}{K}$. Also, $m \geq \frac{b}{\lambda}$. Thus,

$$\left(\frac{c}{b}\right)^6 = \left(\frac{\lambda}{bm}\right)^6 \geq m^{-12} = \exp(-12\ln m) \geq \exp(-Km) = \delta,$$

so Theorem 4.8 applies. For any $\tau \geq 0$, we have

$$\Pr(|X - \mu| \geq \tau) \leq 2\exp\left(\frac{-\tau^2 m}{10\lambda^2\left(1 + \frac{2\tau}{15\lambda}\right)} + m^2\frac{b}{\lambda}\exp\left(-\frac{K}{2}m\right)\exp\left(\frac{\tau bm}{4\lambda^2}\right)\right) + 2m\exp\left(-\frac{K}{2}m\right). \quad (10)$$

Since $\tau \leq \frac{15\lambda}{2}$, we have $1 + \frac{2\tau}{15\lambda} \leq 2$.

Since $\tau \leq \frac{\lambda^2 K}{b}$, we know that $\frac{\tau b m}{4\lambda^2} \leq \frac{K}{4}m$. Also, using $m \geq \frac{b}{\lambda}$, and $\frac{m}{\ln m} > \frac{24}{K}$,

$$m^2 \frac{b}{\lambda} \leq m^3 = \exp(3 \ln m) \leq \exp\left(\frac{K}{8}m\right).$$

Similarly, $m \leq \exp\left(\frac{K}{24}m\right)$. So, by Inequality (10), we get

$$\Pr(|X - \mu| \geq \tau) \leq 2\exp\left(\frac{-\tau^2 m}{20\lambda^2} + \exp\left(-\frac{K}{8}m\right)\right) + 2\exp\left(-\frac{11K}{24}m\right). \tag{11}$$

Now, since $m \geq \lambda\sqrt{40}$ and $\tau \geq 1/m$,

$$\exp\left(-\frac{K}{8}m\right) \leq m^{-3} \leq \frac{1}{40\lambda^2 m} \leq \frac{\tau^2 m}{40\lambda^2},$$

so Inequality (11) implies

$$\Pr(|X - \mu| \geq \tau) \leq 2\exp\left(-\frac{\tau^2 m}{40\lambda^2}\right) + 2\exp\left(-\frac{11K}{24}m\right). \tag{12}$$

Finally, since $\tau \leq 4\lambda\sqrt{K}$,

$$\frac{\tau^2 m}{40\lambda^2} \leq \frac{2K}{5}m < \frac{11K}{24}m,$$

so Inequality (12) gives us

$$\Pr(|X - \mu| \geq \tau) \leq 4\exp\left(-\frac{\tau^2 m}{40\lambda^2}\right).$$

■

5 The case when $\lambda = 0$

In the applications in this thesis, $c = \lambda/m$, and $\delta \to 0$ exponentially in m. So the choice of parameters in Theorems 3.6 and 4.8 is close to optimal. In particular, $\delta \leq (c/b)^6$ for sufficiently large m.

However, in some other settings, Theorems 3.6 and 4.8 are less useful; in particular, they are vacuous when $\lambda = c = 0$. We now prove simpler inequalities which applies in this case. We first prove Theorem 5.3, which is an analog of Theorem 4.8, and then Theorem 5.4, which is an analog of Theorem 3.6.

Note 5.1 For $\omega \in \Omega = \prod_{k=1}^m \Omega_k$, $i \in \{1, \ldots, m\}$, and $v \in \Omega_i$, we let $\omega^{i,v}$ denote the element of Ω obtained by replacing the ith coordinate of ω with v.

Lemma 5.2 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^m \Omega_k$, and let X be a random variable on Ω.*

1. *If X is weakly difference-bounded by $(b, 0, \delta)$, then, for any $\omega, \chi \in \Omega$,*

$$|X(\omega) - X(\chi)| \leq mb.$$

2. *If X is weakly difference-bounded by $(b, 0, \delta)$, then there is some $\chi \in \Omega$ such that*

$$\Pr(X \neq X(\chi)) \leq m\delta.$$

3. *If X is strongly difference-bounded by $(b, 0, \delta)$, then there is some $\chi \in \Omega$ such that*

$$\Pr(X \neq X(\chi)) \leq \frac{m\delta}{2}.$$

Proof: Given any two points $\omega = (\omega_1, \ldots, \omega_m)$ and $\chi = (\chi_1, \ldots, \chi_m)$ in Ω, we define $\psi_i \in \Omega$ as follows: let $\psi_0 = \omega$. For $1 \leq i \leq m$, let

$$\psi_i = (\psi_{i-1})^{i,\chi_i} = (\chi_1, \ldots, \chi_i, \omega_{i+1}, \ldots, \omega_m),$$

so $\psi_m = \chi$.

Proof of Part 1: Suppose X is weakly difference-bounded by $(b, 0, \delta)$. Then, for any i,

$$|X(\psi_{i-1}) - X(\psi_i)| \leq b.$$

Summing over i, we conclude

$$|X(\omega) - X(\chi)| \leq mb.$$

□

Proof of Part 2: Suppose X is weakly difference-bounded by $(b, 0, \delta)$. For any i,

$$\Pr_{\omega,\chi}(X(\psi_{i-1}) \neq X(\psi_i)) \leq \delta.$$

Hence, adding up the probabilities,

$$\Pr_{\omega,\chi}(X(\omega) \neq X(\chi)) \leq m\delta.$$

By a total probability argument, we conclude that there exists some $\chi \in \Omega$ for which $\Pr_\omega(X(\omega) \neq X(\chi)) \leq m\delta$. □

Proof of Part 3: Suppose X is strongly difference-bounded by $(b, 0, \delta)$. Let B denote the bad set of Definition 1.6. For any i,

$$\Pr_{\omega,\chi}(\psi_i \in B) \leq \delta.$$

So, the probability that $\psi_i \in B$ for some odd i is at most $m\delta/2$. If $\psi_i \notin B$ for every odd i, we must have $X(\omega) = X(\chi)$. Therefore,

$$\Pr_{\omega,\chi}(X(\omega) \neq X(\chi)) \leq \frac{m\delta}{2}.$$

By a total probability argument, we conclude that there exists some $\chi \in \Omega$ for which $\Pr_\omega(X(\omega) \neq X(\chi)) \leq m\delta/2$ □

This concludes the lemma. ■

Theorem 5.3 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω which is weakly difference-bounded by $(b, 0, \delta)$. Let $\mu = \mathbf{E}(X)$. Then*

$$\Pr(|X - \mu| > m^2 b\delta) \leq m\delta.$$

Proof: By Part 2 of Lemma 5.2, there is some $\chi \in \Omega$ for which

$$\Pr(|X - X(\chi)| > 0) \leq m\delta.$$

By Part 1 of Lemma 5.2,

$$|X(\omega) - X(\chi)| \leq mb$$

for every $\omega \in \Omega$. We conclude that

$$\mathbf{E}(|X - X(\chi)|) \leq m^2 b\delta,$$

which immediately implies that

$$|\mu - X(\chi)| \leq m^2 b\delta.$$

So,

$$\Pr\left(|X - \mu| > m^2 b\delta\right) \leq \Pr(X \neq X(\chi)) \leq m\delta.$$

■

Theorem 5.4 *Let $\Omega_1, \ldots, \Omega_m$ be probability spaces. Let $\Omega = \prod_{k=1}^{m} \Omega_k$, and let X be a random variable on Ω which is strongly difference-bounded by $(b, 0, \delta)$. Let $\mu = \mathbf{E}(X)$. Then*

$$\Pr\left(|X - \mu| > \frac{m^2 b\delta}{2}\right) \leq \frac{m\delta}{2}.$$

Proof: The proof is similar to that of Theorem 5.3. By Part 3 of Lemma 5.2, there is some $\chi \in \Omega$ for which

$$\Pr(|X - X(\chi)| > 0) \leq \frac{m\delta}{2}.$$

By Part 1 of Lemma 5.2 (we recall from Note 1.8 that strong difference-boundedness implies weak difference-boundedness),

$$|X(\omega) - X(\chi)| \leq mb$$

for every $\omega \in \Omega$. We conclude that

$$\mathbf{E}(|X - X(\chi)|) \leq \frac{m^2 b\delta}{2},$$

which immediately implies that

$$|\mu - X(\chi)| \leq \frac{m^2 b\delta}{2}.$$

So,

$$\Pr\left(|X - \mu| > \frac{m^2 b\delta}{2}\right) \leq \Pr(X \neq X(\chi)) \leq \frac{m\delta}{2}.$$

■

6 Open questions

Question 6.1 Can we prove a continuous version of Theorem 3.3 or Theorem 4.5?

McDiarmid's original inequality, Theorem 3.2, requires a single bound on $|X(\omega) - X(\omega')|$ for all ω, ω'. Our Theorem 4.5 uses two bounds: a large bound b which holds everywhere, and a smaller bound c which holds for almost all choices of ω and ω'.

For any k, we can define a random variable on $\Omega \times \Omega_k$: we let

$$\Delta(\omega, \upsilon) = |X(\omega) - X(\omega')|,$$

where ω' is ω with the kth entry replaced by υ.

McDiarmid's Theorem 3.2 uses $\sup(\Delta)$, and the proof uses Lemma 2.7, which bounds $\mathbf{E}(e^X)$ in terms of $\sup(X)$. Theorem 4.5 is based on bounding Δ almost everywhere, and the proof uses Lemma 2.8, which bounds $\mathbf{E}(e^X)$ given exactly such information about a random variable X.

Can we prove a more general McDiarmid-like inequality, giving a concentration bound in terms of properties of Δ (e.g., $\sup(\Delta), \mathbf{E}(\Delta), \mathrm{Var}(\Delta)$)?

McAllester [15] expresses a similar desire for a variance-based strengthening of Theorem 3.2.

Such a result would have significant implications for the theory of algorithmic stability [5, 13].

Question 6.2 Recently, Talagrand [21] proved a new inequality which can be used to generalize some applications of McDiarmid's Theorem. McDiarmid [19] gives an overview of Talagrand's inequality with some applications. Boucheron, et al. [4] discuss McDiarmid's inequality and Talagrand's inequality, as well as other concentration inequalities, with an emphasis on learning theory applications.

Can we use Talagrand's inequality to simplify the proofs in this paper, or to strengthen the results?

Question 6.3 We prove Theorem 3.6 by first proving Theorem 3.3, and then choosing α_k. We prove Theorem 4.8 by first proving Theorem 4.5, and then choosing θ_k and γ_k. The choices we make for these parameters are based on the assumption that $b = \Theta(1)$, $c = \Theta(1/m)$, and $\delta = \exp(-\Omega(m))$.

Can we choose these parameters more generally? What values of b, c, and δ arise in natural applications, and what choices of α_k, γ_k, and θ_k are best in these situations?

Acknowledgements. I would like to thank Partha Niyogi for introducing me to this area, and for many helpful discussions. I would also like to thank Laci Babai, André Elisseeff, and Pradyut Shah for their comments.

References

[1] K. Azuma. Weighted sums of certain dependent random variables. *Tôhoku Mathematical Journal*, 19(3):357–367, 1967.

[2] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[3] B. Bollobás. The chromatic number of random graphs. *Combinatorica*, 8(1):49–55, 1988.

[4] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16(3):277–292, 2000.

[5] O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In *NIPS*, 2001.

[6] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23(4):493–507, 1952.

[7] D. Freedman. On tail probabilities for martingales. *Annals of Probability*, 3(1):100–118, 1975.

[8] Y. Freund, Y. Mansour, and R. Schapire. Why averaging classifiers can protect against overfitting. In *Workshop on Artificial Intelligence and Statistics*, 2001.

[9] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[10] J. Kahn. Asymptotically good list-colorings. *Journal of Combinatorial Theory, Series A*, 73(1):1–59, 1996.

[11] S. Kutin. Concentration bounds for Good-Turing estimators. Technical report, Department of Computer Science, The University of Chicago, 2002. In preparation.

[12] S. Kutin and P. Niyogi. The interaction of stability and weakness in AdaBoost. Technical Report TR-2001-30, Department of Computer Science, The University of Chicago, 2001.

[13] S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. Technical Report TR-2002-03, Department of Computer Science, The University of Chicago, 2002.

[14] B. Maurey. Construction de suites symétriques. *Comptes Rendus des Séances de l'Académie des Sciences, Paris, Série A: Sciences Mathématiques*, 288(14):679–681, 1979.

[15] D. McAllester. Two variance-based concentration inequalities with applications. Under review; available online, 2001.

[16] D. McAllester and R. Schapire. On the convergence rate of Good-turing estimators. In *COLT*, 2000.

[17] D. McAllester and R. Schapire. Learning theory and language modeling. In *IJCAI*, 2001.

[18] C. McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[19] C. McDiarmid. Concentration. In *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, Berlin, 1998.

[20] W. Steiger. A best possible Kolmogoroff-type inequality for martingales and a characteristic property. *Annals of Mathematical Statistics*, 40(3):764–769, 1969.

[21] M. Talagrand. Concentration of measure and isoperimetric inequalities in product spaces. *Publications Math'ematiques, Institut des Hautes Études Scientifiques*, 81:73–205, 1995.