The Median Method

Emin Karayel

March 17, 2025

Abstract

The median method is an amplification result for randomized approximation algorithms described in [1]. Given an algorithm whose result is in a desired interval with a probability larger than $\frac{1}{2}$, it is possible to improve the success probability, by running the algorithm multiple times independently and using the median. In contrast to using the mean, the amplification of the success probability grows exponentially with the number of independent runs.

This entry contains a formalization of the underlying theorem: Given a sequence of n independent random variables, which are in a desired interval with a probability $\frac{1}{2} + \alpha$. Then their median will be in the desired interval with a probability of $1 - \exp(-2\alpha^2 n)$. In particular, the success probability approaches 1 exponentially with the number of variables.

In addition to that, this entry also contains a proof that order-statistics of Borel-measurable random variables are themselves measurable and that generalized intervals in linearly ordered Borel-spaces are measurable.

Contents

1 Intervals are Borel measurable

theory *Median*
imports
HOL−Probability.Probability
HOL−Library.Multiset
Universal-Hash-Families.Universal-Hash-Families-More-Independent-Families

begin

This section contains a proof that intervals are Borel measurable, where an interval is defined as a convex subset of linearly ordered space, more precisely, a set is an interval, if for each triple of points $x < y < z$: If x and z are in the set so is y. This includes ordinary intervals like $\{a..b\}$, $\{a<..<b\}$ but also for example $\{x::rat.\ x * x < (2::rat)\}$ which cannot be expressed in the standard notation.

In the *HOL−Analysis.Borel-Space* there are proofs for the measurability of each specific type of interval, but those unfortunately do not help if we want to express the result about the median bound for arbitrary types of intervals.

definition *interval* :: ($'a$ :: *linorder*) *set* ⇒ *bool* **where**
  interval $I = (\forall x\ y\ z.\ x \in I \longrightarrow z \in I \longrightarrow x \leq y \longrightarrow y \leq z \longrightarrow y \in I)$

definition *up-ray* :: ($'a$ :: *linorder*) *set* ⇒ *bool* **where**
  up-ray $I = (\forall x\ y.\ x \in I \longrightarrow x \leq y \longrightarrow y \in I)$

lemma *up-ray-borel*:
  assumes *up-ray* (I :: (($'a$ :: *linorder-topology*) *set*))
  shows $I \in$ *borel*
⟨*proof*⟩

definition *down-ray* :: ($'a$ :: *linorder*) *set* ⇒ *bool* **where**
  down-ray $I = (\forall x\ y.\ y \in I \longrightarrow x \leq y \longrightarrow x \in I)$

lemma *down-ray-borel*:
  assumes *down-ray* (I :: (($'a$ :: *linorder-topology*) *set*))
  shows $I \in$ *borel*
⟨*proof*⟩

Main result of this section:

lemma *interval-borel*:
  assumes *interval* (I :: (($'a$ :: *linorder-topology*) *set*))
  shows $I \in$ *borel*
⟨*proof*⟩

2 Order statistics are Borel measurable

This section contains a proof that order statistics of Borel measurable random variables are themselves Borel measurable.

The proof relies on the existence of branch-free comparison-sort algorithms. Given a sequence length these algorithms perform compare-swap operations on predefined pairs of positions. In particular the result of a comparison does not affect future operations. An example for a branch-free comparison sort algorithm is shell-sort and also bubble-sort without the early exit.

The advantage of using such a comparison-sort algorithm is that it can be lifted to work on random variables, where the result of a comparison-swap operation on two random variables X and Y can be represented as the expressions $\lambda\omega.\ min\ (X\ \omega)\ (Y\ \omega)$ and $\lambda\omega.\ max\ (X\ \omega)\ (Y\ \omega)$.

Because taking the point-wise minimum (resp. maximum) of two random variables is still Borel measurable, and because the entire sorting operation can be represented using such compare-swap operations, we can show that all order statistics are Borel measuable.

fun *sort-primitive* **where**
  sort-primitive i j f k = (*if k* = *i then min* (*f i*) (*f j*) *else* (*if k* = *j then max* (*f i*) (*f j*) *else f k*))

fun *sort-map* **where**
  sort-map f n = *fold id* [*sort-primitive j i*. *i* <− [*0*..<*n*], *j* <− [*0*..<*i*]] *f*

lemma *sort-map-ind*:
  sort-map f (*Suc n*) = *fold id* [*sort-primitive j n*. *j* <− [*0*..<*n*]] (*sort-map f n*)
  ⟨*proof*⟩

lemma *sort-map-strict-mono*:
  fixes *f* :: *nat* ⇒ ′*b* :: *linorder*
  shows $j < n \Longrightarrow i < j \Longrightarrow$ *sort-map f n i* ≤ *sort-map f n j*
⟨*proof*⟩

lemma *sort-map-mono*:
  fixes *f* :: *nat* ⇒ ′*b* :: *linorder*
  shows $j < n \Longrightarrow i \le j \Longrightarrow$ *sort-map f n i* ≤ *sort-map f n j*
  ⟨*proof*⟩

lemma *sort-map-perm*:
  fixes *f* :: *nat* ⇒ ′*b* :: *linorder*
  shows *image-mset* (*sort-map f n*) (*mset* [*0*..<*n*]) = *image-mset f* (*mset* [*0*..<*n*])
⟨*proof*⟩

lemma *list-eq-iff*:
  assumes *mset xs* = *mset ys*
  assumes *sorted xs*
  assumes *sorted ys*
  shows *xs* = *ys*
  ⟨*proof*⟩

lemma *sort-map-eq-sort*:
  fixes *f* :: *nat* ⇒ (′*b* :: *linorder*)
  shows *map* (*sort-map f n*) [*0*..<*n*] = *sort* (*map f* [*0*..<*n*]) (**is** *?A* = *?B*)
⟨*proof*⟩

lemma *order-statistics-measurable-aux*:
  fixes *X* :: *nat* ⇒ ′*a* ⇒ (′*b* :: {*linorder-topology*, *second-countable-topology*})

assumes $n \geq 1$
assumes $j < n$
assumes $\bigwedge i.\ i < n \Longrightarrow X\ i \in \mathit{measurable}\ M\ \mathit{borel}$
shows $(\lambda x.\ (\mathit{sort\text{-}map}\ (\lambda i.\ X\ i\ x)\ n)\ j) \in \mathit{measurable}\ M\ \mathit{borel}$
⟨*proof*⟩

Main results of this section:

lemma *order-statistics-measurable*:
fixes $X :: \mathit{nat} \Rightarrow 'a \Rightarrow ('b :: \{\mathit{linorder\text{-}topology},\ \mathit{second\text{-}countable\text{-}topology}\})$
assumes $n \geq 1$
assumes $j < n$
assumes $\bigwedge i.\ i < n \Longrightarrow X\ i \in \mathit{measurable}\ M\ \mathit{borel}$
shows $(\lambda x.\ (\mathit{sort}\ (\mathit{map}\ (\lambda i.\ X\ i\ x)\ [0..{<}n]))\ !\ j) \in \mathit{measurable}\ M\ \mathit{borel}$
⟨*proof*⟩

definition $\mathit{median} :: \mathit{nat} \Rightarrow (\mathit{nat} \Rightarrow ('a :: \mathit{linorder})) \Rightarrow 'a$ **where**
$\mathit{median}\ n\ f = \mathit{sort}\ (\mathit{map}\ f\ [0..{<}n])\ !\ (n\ \mathit{div}\ 2)$

lemma *median-measurable*:
fixes $X :: \mathit{nat} \Rightarrow 'a \Rightarrow ('b :: \{\mathit{linorder\text{-}topology},\ \mathit{second\text{-}countable\text{-}topology}\})$
assumes $n \geq 1$
assumes $\bigwedge i.\ i < n \Longrightarrow X\ i \in \mathit{measurable}\ M\ \mathit{borel}$
shows $(\lambda x.\ \mathit{median}\ n\ (\lambda i.\ X\ i\ x)) \in \mathit{measurable}\ M\ \mathit{borel}$
⟨*proof*⟩

3 The Median Method

This section contains the proof for the probability that the median of independent random variables will be in an interval with high probability if the individual variables are in the same interval with probability larger than $\frac{1}{2}$.
The proof starts with the elementary observation that the median of a seqeuence with n elements is in an interval I if at least half of them are in I. This works because after sorting the sequence the elements that will be in the interval must necessarily form a consecutive subsequence, if its length is larger than $\frac{n}{2}$ the median must be in it.
The remainder follows the proof in [1, §2.1] using the Hoeffding inequality to estimate the probability that at least half of the sequence elements will be in the interval I.

lemma *interval-rule*:
assumes *interval* I
assumes $a \leq x\ x \leq b$
assumes $a \in I$
assumes $b \in I$
shows $x \in I$
⟨*proof*⟩

lemma *sorted-int*:
assumes *interval I*
assumes *sorted xs*
assumes $k <$ *length xs* $i \leq j$ $j \leq k$
assumes *xs* ! $i \in I$ *xs* ! $k \in I$
shows *xs* ! $j \in I$
⟨*proof*⟩

lemma *mid-in-interval*:
assumes *2∗length* (*filter* ($\lambda x.\ x \in I$) *xs*) $>$ *length xs*
assumes *interval I*
assumes *sorted xs*
shows *xs* ! (*length xs div 2*) $\in I$
⟨*proof*⟩

lemma *median-est*:
assumes *interval I*
assumes *2∗card* $\{k.\ k < n \land f\ k \in I\} > n$
shows *median n f* $\in I$
⟨*proof*⟩

lemma *median-est-rev*:
assumes *interval I*
assumes *median n f* $\notin I$
shows *2∗card* $\{k.\ k < n \land f\ k \notin I\} \geq n$
⟨*proof*⟩

lemma *prod-pmf-bernoulli-mono*:
assumes *finite I*
assumes $\bigwedge i.\ i \in I \Longrightarrow 0 \leq f\ i \land f\ i \leq g\ i \land g\ i \leq 1$
assumes $\bigwedge x\ y.\ x \in A \Longrightarrow (\forall\, i \in I.\ x\ i \leq y\ i) \Longrightarrow y \in A$
shows *measure* (*Pi-pmf I d* (*bernoulli-pmf* $\circ$ *f*)) *A* $\leq$ *measure* (*Pi-pmf I d* (*bernoulli-pmf* $\circ$ *g*)) *A*
(**is** *?L* $\leq$ *?R*)
⟨*proof*⟩

lemma *discrete-measure-eqI*:
assumes *sets M* = *count-space UNIV*
assumes *sets N* = *count-space UNIV*
assumes *countable* Ω
assumes $\bigwedge x.\ x \in \Omega \Longrightarrow$ *emeasure M* $\{x\}$ = *emeasure N* $\{x\} \land$ *emeasure M* $\{x\} \neq \infty$
assumes *AE x in M*. $x \in \Omega$
assumes *AE x in N*. $x \in \Omega$
shows $M = N$
⟨*proof*⟩

Main results of this section:

The next theorem establishes a bound for the probability of the median of independent random variables using the binomial distribution. In a follow-up step, we will establish tail bounds for the binomial distribution and corresponding median bounds.

This two-step strategy was suggested by Yong Kiam Tan. In a previous version, I only had verified the exponential tail bound (see theorem `median_bound` below).

theorem (**in** *prob-space*) *median-bound-raw*:
  fixes I :: ($'b$ :: {*linorder-topology*, *second-countable-topology*}) *set*
  assumes $n > 0$ $p \geq 0$
  assumes *interval* I
  assumes *indep-vars* (λ-. *borel*) X $\{0..<n\}$
  assumes $\bigwedge i.\ i < n \Longrightarrow \mathcal{P}(\omega\ in\ M.\ X\ i\ \omega \in I) \geq p$
  shows $\mathcal{P}(\omega\ in\ M.\ median\ n\ (\lambda i.\ X\ i\ \omega) \in I) \geq 1 - measure\ (binomial\text{-}pmf\ n\ p)$ $\{..n\ div\ 2\}$
    (**is** $?L \geq ?R$)
⟨*proof*⟩

Cumulative distribution of the binomial distribution (contributed by Yong Kiam Tan):

lemma *prob-binomial-pmf-upto*:
  assumes $0 \leq p$ $p \leq 1$
  shows *measure-pmf.prob* (*binomial-pmf* n p) $\{..m\}$ =
    $sum\ (\lambda i.\ real\ (n\ choose\ i) * p\hat{\ }i * (1 - p)\ \hat{\ }(n-i))\ \{0..m\}$
  ⟨*proof*⟩

A tail bound for the binomial distribution using Hoeffding's inequality:

lemma *binomial-pmf-tail*:
  assumes $p \in \{0..1\}$ $real\ k \leq real\ n * p$
  shows $measure\ (binomial\text{-}pmf\ n\ p)\ \{..k\} \leq exp\ (-\ 2 * real\ n * (p - real\ k\ /\ n)\hat{\ }2)$
    (**is** $?L \leq ?R$)
⟨*proof*⟩

theorem (**in** *prob-space*) *median-bound*:
  fixes n :: *nat*
  fixes I :: ($'b$ :: {*linorder-topology*, *second-countable-topology*}) *set*
  assumes *interval* I
  assumes $\alpha > 0$
  assumes $\varepsilon \in \{0<..<1\}$
  assumes *indep-vars* (λ-. *borel*) X $\{0..<n\}$
  assumes $n \geq -\ ln\ \varepsilon\ /\ (2 * \alpha^2)$
  assumes $\bigwedge i.\ i < n \Longrightarrow \mathcal{P}(\omega\ in\ M.\ X\ i\ \omega \in I) \geq 1/2+\alpha$
  shows $\mathcal{P}(\omega\ in\ M.\ median\ n\ (\lambda i.\ X\ i\ \omega) \in I) \geq 1-\varepsilon$
⟨*proof*⟩

This is a specialization of the above to closed real intervals.

corollary (**in** *prob-space*) *median-bound-1*:

assumes $\alpha > 0$
assumes $\varepsilon \in \{0 <..< 1\}$
assumes *indep-vars* (λ-. *borel*) X $\{0..<n\}$
assumes $n \geq - \ln \varepsilon \ / \ (2 * \alpha^2)$
assumes $\forall i \in \{0..<n\}.\ \mathcal{P}(\omega\ in\ M.\ X\ i\ \omega \in (\{a..b\} :: real\ set)) \geq 1/2 + \alpha$
shows $\mathcal{P}(\omega\ in\ M.\ median\ n\ (\lambda i.\ X\ i\ \omega) \in \{a..b\}) \geq 1 - \varepsilon$
⟨*proof*⟩

This is a specialization of the above, where $\alpha = \frac{1}{6}$ and the interval is described using a mid point μ and radius δ. The choice of $\alpha = \frac{1}{6}$ implies a success probability per random variable of $\frac{2}{3}$. It is a commonly chosen success probability for Monte-Carlo algorithms (cf. [2, §4] or [3, §1]).

corollary (**in** *prob-space*) *median-bound-2*:
fixes $\mu\ \delta$ :: *real*
assumes $\varepsilon \in \{0 <..< 1\}$
assumes *indep-vars* (λ-. *borel*) X $\{0..<n\}$
assumes $n \geq -18 * \ln \varepsilon$
assumes $\bigwedge i.\ i < n \Longrightarrow \mathcal{P}(\omega\ in\ M.\ abs\ (X\ i\ \omega - \mu) > \delta) \leq 1/3$
shows $\mathcal{P}(\omega\ in\ M.\ abs\ (median\ n\ (\lambda i.\ X\ i\ \omega) - \mu) \leq \delta) \geq 1 - \varepsilon$
⟨*proof*⟩

4 Some additional results about the median

lemma *sorted-mono-map*:
assumes *sorted xs*
assumes *mono f*
shows *sorted* (*map f xs*)
⟨*proof*⟩

This could be added to *HOL.List*:

lemma *map-sort*:
assumes *mono f*
shows *sort* (*map f xs*) = *map f* (*sort xs*)
⟨*proof*⟩

lemma *median-cong*:
assumes $\bigwedge i.\ i < n \Longrightarrow f\ i = g\ i$
shows *median n f* = *median n g*
⟨*proof*⟩

lemma *median-restrict*:
$median\ n\ (\lambda i \in \{0..<n\}.f\ i) = median\ n\ f$
⟨*proof*⟩

lemma *median-commute-mono*:
assumes $n > 0$
assumes *mono g*
shows $g\ (median\ n\ f) = median\ n\ (g \circ f)$

$\langle proof \rangle$

lemma *median-rat*:
assumes $n > 0$
shows $real\text{-}of\text{-}rat\ (median\ n\ f) = median\ n\ (\lambda i.\ real\text{-}of\text{-}rat\ (f\ i))$
$\langle proof \rangle$

lemma *median-const*:
assumes $k > 0$
shows $median\ k\ (\lambda i \in \{0..<k\}.\ a) = a$
$\langle proof \rangle$

end

References

[1] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

[2] Z. Bar-Yossef, T. S. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream. In *Randomization and Approximation Techniques in Computer Science*, pages 1–10. Springer Berlin Heidelberg, 2002.

[3] D. M. Kane, J. Nelson, and D. P. Woodruff. An optimal algorithm for the distinct elements problem. In *Proceedings of the Twenty-Ninth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '10, pages 41–52, New York, 2010.