A Broad Investigation of Median-Based Concentration Inequalities

Nathan Tung

Motivation: Given a function $f(X_1, X_2, \dots, X_n) : \mathbb{X} \to \mathbb{R}$ of independent random variables with some sort of regularity, an extremely important consideration is quantifying its concentration about a central value of interest. Bounding the probability of large random fluctuations is vital to the study of modern learning theory, statistical mechanics, and stochastic processes. Current research tends to attack this problem by bounding $\mathbb{P}\left(|f - \mathbb{E}_f| > t\right)$, where $\mathbb{E}_f = \mathbb{E}[f(X_1, \dots, X_n)]$. The sharpest bounds of this type have been developed through incredibly sophisticated methods to show high-dimensional (functions of many random variables) tail behavior of f, with much of the work being done to achieve efficient "tensorization" as desired tail behavior in one dimension (for a single random variable) is relatively easy to show.

Another approach, however, looks at the probabilities of sets instead of the tail behavior of functions directly and is known as the isoperimetric approach. The probabilities of sets in high-dimensional product spaces can be translated to random variables and this approach tends to gives its tightest bounds on $\mathbb{P}\left(|f - \mathbb{M}_f| > t\right)$, where $\mathbb{M}_f$ is the median of f defined as

$$\mathbb{M}_f = \inf_{A \subseteq \mathcal{X}^n, \mathbb{P}(A) > 1/2} f(A)$$

Talagrand's convex distance inequality is perhaps the most widely used isoperimetric concentration inequality. Given $X_1, \dots, X_n$ taking values in $[0, 1]$ and 1-Lipschitz function $f(X_1, \dots, X_n) : \mathbb{X} \to \mathbb{R} \ni |f(x) - f(y)| \leq d(x, y)$ on metric space $(\mathbb{X}, d)$ we have that [2]:

$$\mathbb{P}(|f - \mathbb{M}_f| > t) \leq 4e^{-t^2/4}$$

Given the numerous ways to show concentration, all relying on subtly different regularity conditions and yielding subtly different results, I believe the relationship between concentration about $\mathbb{E}_f$ and $\mathbb{M}_f$ deserves further exploration. It is worth asking in what sort of problems concentration about $\mathbb{M}_f$ may be desirable, whether advanced methods (entropy, transportation) used for $\mathbb{E}_f$ can be applied to $\mathbb{M}_f$, how such approaches compare to isoperimetric ones, and the merit of considering $\mathbb{M}_f$ in the context of Rademacher Complexity. This proposal raises numerous questions and I cannot hope to address them all. I will aim to narrow them down as I investigate further.

Aim I - Adapt methods for median concentration: The Cramer-Chernoff method gives concentration for single random variables and importantly results in useful formalizations of several function classes by their tail behavior. With $\psi_Z(\lambda) = \mathbb{E}[e^{\lambda Z}]$ the log moment-generating function (MGF) of a random variable Z, let $\psi_Z^*(t) = \sup_{\lambda \geq 0} (\lambda t - \psi_Z(\lambda))$ be the Fenchel-Legendre dual of Z. Then using Chernoff's bound one can obtain that

$$\mathbb{P}(Z > t) \leq e^{-\psi_Z^*(t)} \tag{1}$$

and a bound for the other tail can be obtained by considering $-Z$. Notably the distribution of Z is not used to prove the above bound: all that is needed is an upper bound on $\psi_Z(\lambda)$. By examining the log MGF of common distributions and setting $Z = X - \mathbb{E}[X]$ we can define classes of random variables characterized by their tail behavior. For example, given a random variable of interest X and letting $Z = X - \mathbb{E}[X]$, if we can show that $\psi_Z(\lambda) \leq \psi_G(\lambda) = \frac{\lambda^2 \sigma^2}{2}$ where $\psi_G(\lambda)$ is the log MGF of $G \sim N(0, \sigma^2)$ then Z respects (1) with $\psi_Z^*(t) \geq \psi_G^*(t) = \frac{t^2}{2\sigma^2}$ and we have a bound on the probability of large deviations of X. In this case X is said to be σ^2-*subgaussian*. The same process can be carried out to define *subpoisson* and *subgamma* functions [1].

Higher-dimensional concentration inequalities can then be obtained starting from a variety of different regularity conditions. For example, for a function $f(X_1, ..., X_n)$ define

$$D_i f(x) := \sup_z f(x_1, ..., x_{i-1}, z, x_{i+1}, ..., x_n) - \inf_z f(x_1, ..., x_{i-1}, z, x_{i+1}, ..., x_n)$$

Then the so-called "entropy method" uses the result that

$$\text{Ent}\left[e^{\lambda X}\right] \leq \frac{\lambda^2\sigma^2}{2}\mathbb{E}\left[e^{\lambda X}\right], \forall\lambda \geq 0 \implies X\ \sigma^2\text{-}subgaussian \tag{2}$$

and the tensorization of entropy to obtain that f is subgaussian with $\sigma^2 = 2\left\|\sum_{i=1}^n |D_i f|^2\right\|_\infty$ [5].

If one desires to examine the concentration of L-Lipschitz functions

$$f(X_1, ..., X_n) : \mathbb{X} \to \mathbb{R} \ni |f(x) - f(y)| \leq Ld(x, y)$$

in a metric space $(\mathbb{X}, d)$, the "transportation method" then uses the result that for all $f \in \text{Lip}(\mathbb{X})$

$$W_1(\nu, \mu) \leq \sqrt{2\sigma^2 D(\nu\|\mu)} \iff f(X)\ \sigma^2\text{-}subgaussian \tag{3}$$

for all ν, where μ, ν are probability measures on $(\mathbb{X}, d)$, $X \sim \mu$, $W_1(\mu, \nu) := \sup_{f\in\text{Lip}(\mathbb{X})} |\int f d\mu - \int f d\nu|$ is the Wasserstein Distance, and $D(\nu\|\mu)$ is relative entropy, also known as the Kullbeck-Leibler divergence. Interestingly there is a direct connection between (3) and isoperimetry in that we can evaluate the Wasserstein distance and relative entropy at μ conditioned on $A \subseteq \mathbb{X}$ and $B = \mathbb{X}\backslash A^\varepsilon$, where A^ε is the canonical epsilon-fattening of A in isoperimetric problems, to directly obtain $\mu(A^\varepsilon) \geq 1 - 2e^{-\varepsilon^2/8\sigma^2}$ $\forall\varepsilon \geq 0$, $\mu(A) \geq \frac{1}{2}$ [5].

While I omit the details, (2) and (3) work by simply reformulating the log MGF bounding condition of Cramer-Chernoff in a way that "tensorizes" to high dimensions [5]. This fact leads me to hope that such a framework can be reworked to give bounds involving $\mathbb{M}_f$ if we can upper bound the log MGF of $Z = X - \mathbb{M}[X]$ by the log MGF of some distribution and then adapt the above methods for tensorization. First, the Cramer-Chernoff method would have to be adapted for $Z = X - \mathbb{M}[X]$. There is some subtlety here as in practice the calculation of the Legendre dual is simplified by taking advantage of the fact that $Z = X - \mathbb{E}[X]$ has zero expectation [1]. Changing Z to be median-centered will require reworking the log MGF upper bounds and recomputation of the duals for all three classes, possibly with something like median-parametrized distributions. While *subgaussian* has received the most attention, the weaker *subpoisson* and *subgamma* are interesting because of their ability to capture single-tail bounds. This is important because in many applications we only have one-sided regularity conditions and often single-tail bounds are all that is needed. Once analagous function classes describing median concentration are derived I hope to extend the entropy/transportation methods to tensorize them to higher dimensions.

A possible simpler approach for arriving at median concentration is simply by bounding $|\mathbb{M}_f - \mathbb{E}_f|$ and then directly converting the sharpest existing bounds from $\mathbb{E}_f$ to $\mathbb{M}_f$. While this has been done it tends to result in suboptimal bounds and my hopes are that the above approach will prove more fruitful.

Aim II - Find applications in computer science and compute medians: It is helpful to be able to find $\mathbb{M}_f$ or even bound $|\mathbb{M}_f - \mathbb{E}_f|$ for distributions of interest in order to find an alternative condition to $\psi(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$ and to compare any bounds reached in aim I to existing isoperimetric median bounds. For example, it can be shown that for any random variable with mean μ and variance σ^2, the median must lie in the interval $[\mu - \sigma, \mu + \sigma]$. [3] proposes a set of theorems that can be used to compute more sophisticated median estimates. For example, with $X_n = x_1 + x_2 + ... + x_n$ a sum of independent

heterogeneous Bernoulli Trials it is shown that $\mathbb{P}(X_n > \mathbb{E}[X_n]) < \frac{1}{2} < \mathbb{P}(X_n \geq \mathbb{E}[X_n])$. Since the Poisson distribution can be attained as a limit of Bernoulli Trials, its median can be attained using this. This framework is also used to analyze the interpolation search algorithm and Polya's Urn-type problems. I hope to use insights from [3] to find applications where median concentration inequalities are especially useful. One example, presented in [5], are norms of n-dimensional centered Gaussian vectors where it is easier to consider the median of the norm than the expectation in analysis. Of particular interest are functions on the hypercube as the classical vertex and edge isoperimetric problems are closely related to concentration and it seems especially median concentration. I hope to consider a variety of binomial and Rademacher functions on the hypercube as well as algorithms in computer science that heavily rely on medians.

Aim III - Applications in statistical learning and Rademacher Complexity: Concentration inequalities are key to some of the most fundamental questions in machine learning, namely hypothesis class complexity and understanding the sample size required for approximation guarantees. I am hopeful that median-based bounds could reformulate the way we think about some of these long-standing questions. As simple concentration inequalities are used heavily throughout the theory of PAC learning, I suspect there are applications of median-bounds in the theory of ε-nets, ε-samples, and uniform convergence as in [6]. Another possible, albeit radical and ambitious, application is in new approaches to the Rademacher average. The Rademacher average captures the expected error in estimating the expectation of any function f in family $\mathcal{F}$ and is defined as

$$R_m(\mathcal{F}) = \mathbb{E}_{S\sim\mathcal{D}}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right]$$

where $S = (z_1, \ldots, z_m)$ is an m-sample from distribution $\mathcal{D}$ and $\sigma = (\sigma_1, \ldots, \sigma_m)$ is a sequence of iid Rademacher variables with $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$. It can then be derived that the Rademacher average bounds the expected error as [4]

$$\mathbb{E}_{S\sim\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m}f(z_i)\right)\right] \leq 2R_m(\mathcal{F}) \tag{4}$$

If an alternative function to $R_m(\mathcal{F})$, say $S_m(\mathcal{F})$, could be found such that

$$\mathbb{M}_{S\sim\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}_{\mathcal{D}}[f(z)] - \frac{1}{m}\sum_{i=1}^{m}f(z_i)\right)\right] \lesssim S_m(\mathcal{F}) \tag{5}$$

then one could use median-bounds from aim I (or isoperimetry) to attain concentration of the supremum deviation about $S_m(\mathcal{F})$. This would however require reworking the proof of the symmetrization inequality (4) in [4] which relies heavily on properties of expectation. The desired concentration result would also have to be reproved, addressing case by case the cardinality of $\mathcal{F}$, and bounds/approximations of $S_m(\mathcal{F})$ would have to be found if this theory is to be useful. As an alternative that less directly uses results from aim I but holds more promise in exploring new areas of learning theory one could consider $T_m(\mathcal{F})$ such that

$$\mathbb{E}_{S\sim\mathcal{D}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{M}_{\mathcal{D}}[f(z)] - \hat{\mathbb{M}}_m[f(z_i)]\right)\right] \lesssim T_m(\mathcal{F}) \tag{6}$$

where $\hat{\mathbb{M}}_m[f(z)]$ is an unbiased estimator of $\mathbb{M}_{\mathcal{D}}[f(z)]$. (5) and (6) are very different questions. The former represents a new technique for arriving at concentration in estimating the expectation of $f \in \mathcal{F}$ while the latter is an entirely different notion of complexity.

References

[1] S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities - a nonasymptotic theory of independence. In *Concentration Inequalities*, 2013.

[2] D. Gamarnik. Concentration inequalities and applications: Advanced stochastic processes lecture notes massachusetts institute of technology. https://ocw.mit.edu/courses/sloan-school-of-management/15-070j-advanced-stochastic-processes-fall-2013/lecture-notes/MIT15_070JF13_Lec13.pdf.

[3] A. Siegel. Median bounds and their application. https://cs.nyu.edu/~siegel/median.pdf.

[4] E. Upfal. Rademacher complexity: Csci 2540 lecture notes brown university. http://cs.brown.edu/courses/cs155/slides/2021/class17-18.pdf.

[5] R. van Handel. Probability in high dimension: Apc 550 lecture notes princeton university. https://web.math.princeton.edu/~rvan/APC550.pdf.

[6] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities theory of probability & its applications, 16 (2): 264–280, 1971.