Uniform Convergence, Fair Machine Learning, and Dependent Statistical Estimation;

or My Work in Three Inequalities

Cyrus Cousins

August 2021

Here I present a loosely technical overview of the most significant mathematical ideas in my work, as summarized by three simple inequalities, alongside their broad implications and context. Such a concise presentation is necessarily reductionist, and similarities are emphasized over differences, but the overall character of the work shines through. Crucial context for these inequalities is that each involves a novel or previously underappreciated or overlooked *statistically stable* (easily estimated) quantity, which relates to some well-understood critical measure of the error of a sampling process. For simplicity, I present only *expectation bounds*, but each inequality can be converted to a similar high-probability bound.

A key theme in my work is removing unrealistic or limiting assumptions of *a priori* knowledge on various random processes, showing whenever possible that sample-estimates of key quantities suffice. In particular, via highly precise and technical concentration of measure analysis, I show that the relevant quantities in each of the three inequalities can be estimated from the available data, with modest constant-factor cost, and little or no surplus asymptotic cost. Much of the technical challenge thus revolves around showing rigorous finite-sample concentration guarantees for the relevant quantities, in particular matching lower and upper bounds whenever possible. For the purposes of comparing and understanding such results, asymptotic analyses often suffice, but in practical settings, finite sample analyses, with fully specified (and ideally well-optimized) constants are necessary, and thus are strongly pursued in my work.

However, this overview focuses on the three expectation inequalities themselves, explaining their fundamental utility, and how they can be used to characterize and analyze various estimation problems. The reader should come away with an understanding of why they are important, and ideally a curiosity as to how to compute or bound the relevant quantities from finite samples. The interested reader is also invited to consult my thesis proposal [4] for a brief introduction, or my dissertation [6] for a thoroughly detailed presentation.

1. The Empirically Centralized Rademacher Average [8]

A generic method for deriving upper and lower bounds for the error of simultaneous mean estimation across arbitrary function families.

Suppose *function family* $\mathcal{F} \subseteq \mathcal{X} \to \mathbb{R}$, *i.i.d. data sample* $\boldsymbol{x} \sim \mathcal{D}^m$, and *i.i.d. Rademacher sequence* $\boldsymbol{\sigma} \sim \mathcal{U}^m(\pm 1)$ (uniform on ± 1). I define the *empirically centralized empirical Rademacher average* as

$$\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}], \boldsymbol{x})\big] \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i \Big(f(\boldsymbol{x}_i) - \hat{\mathbb{E}}_{\boldsymbol{x}}[f] \Big)\right] \ ,$$

which intuitively measures the ability of $\mathcal{F}$ to *spuriously correlate* with *noise* ($\boldsymbol{\sigma}$). More precisely, we have

$$\frac{1}{2} \cdot \frac{\mathbb{E}_{\boldsymbol{x}}\big[\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}], \boldsymbol{x})\big]}{1 + \sqrt{\frac{8/\pi}{m - \frac{1}{2}}}} \leq \mathop{\mathbb{E}}_{\boldsymbol{x}}\Bigg[\underbrace{\sup_{f \in \mathcal{F}} \Big| \hat{\mathbb{E}}_{\boldsymbol{x}}[f] - \mathop{\mathbb{E}}_{\mathcal{D}}[f] \Big|}_{\textsc{Supremum Deviation}}\Bigg] \leq 2 \cdot \frac{\mathbb{E}_{\boldsymbol{x}}\big[\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}], \boldsymbol{x})\big]}{1 - \sqrt{\frac{8/\pi}{m - \frac{1}{2}}}} \ , \tag{1}$$

which further implies the asymptotic relationship

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\sup_{f\in\mathcal{F}}\left|\hat{\mathbb{E}}_{\boldsymbol{x}}[f]-\mathop{\mathbb{E}}_{\mathcal{D}}[f]\right|\right]\asymp\mathop{\mathbb{E}}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\mathcal{F}-\hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}],\boldsymbol{x})\right]\ .$$

This has immediate statistical applications in simultaneously estimating means across an arbitrary family of functions $\mathcal{F}$, and is also relevant in supervised and unsupervised machine learning settings, wherein $\mathcal{F}$ is taken to be the loss of every function in some model class, and $\mathbb{E}_{\boldsymbol{x}}\big[\hat{\mathfrak{R}}_m(\mathcal{F}-\hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}],\boldsymbol{x})\big]$ is used to bound the *generalization gap* (degree of overfitting) of the learned function.

For context, I note that the (non-empirically) centralized Rademacher average $\mathbb{E}_{\boldsymbol{x}}\big[\hat{\mathfrak{R}}_m(\mathcal{F}-\mathbb{E}_{\mathcal{D}}[\mathcal{F}],\boldsymbol{x})\big]$ is also known to control the supremum deviation, i.e., it admits bounds similar to (1). However, although it is quite valuable as an analytical tool, the centralized Rademacher average can be difficult to work with given only a sample, since computing or estimating $\hat{\mathfrak{R}}_m(\mathcal{F}-\mathbb{E}_{\mathcal{D}}[\mathcal{F}],\boldsymbol{x})$ requires knowledge of *true expectations* $\mathbb{E}_{\mathcal{D}}[\mathcal{F}]$ to evaluate, rather than just the *sample means* $\hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}]$. One convenient application of empirical centralization is that $\hat{\mathfrak{R}}_m(\mathcal{F}-\hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}],\boldsymbol{x})$ removes this dependency, and it can easily be estimated via the Monte Carlo method, wherein we sample to estimate the expectation over $\boldsymbol{\sigma}$ (Rademacher sequences), and simply compute or upper-bound the supremum given the sampled $\boldsymbol{\sigma}$.

2. The Power Mean (Welfare and Fair Machine Learning) [5]

Fairly quantifying wellbeing and illbeing within populations, with statistical and computational guarantees for fair machine learning applications.

Given g groups, *sentiment vector* $\mathcal{S}\in\mathbb{R}^g_{0+}$, and *probability measure* $\boldsymbol{w}\in\mathbb{R}^g_{0+}$, we define the (weighted) *p-power mean* for $p\in(0,\infty)$ as

$$\mathrm{M}_p(\mathcal{S};\boldsymbol{w})\doteq\sqrt[p]{\sum_{i=1}^{g}\boldsymbol{w}_i\mathcal{S}_i^p}\ .$$

The power mean $\mathrm{M}_p(\cdot;\boldsymbol{w})$ over *per-group risk values* is an objective in fair machine learning tasks; $p=1$ yields a *utilitarian* or *weighted average* optimal model, taking $p\to\infty$ converges to the *minimax*, *robust*, or *egalitarian* optimal model, and $1<p<\infty$ smoothly interpolate between these cases. Choosing p allows the model designer to *explicitly make decisions* about how the learner should make fairness-sensitive tradeoffs within the population. This parallels the *social planner's problem*[1] of welfare economics, in which one allocates resources and creates policies to optimize wellbeing within a society.

When learning to optimize power means, there are two main sources of bias (generalization error). First, as in standard risk (expected loss) minimization, there is *selection bias* over the function family $\mathcal{F}$, which we again control with Rademacher averages. The second issue is more subtle; since power means are *nonlinear functions* of expectations, no simple *unbiased estimator* exists (as with, e.g., the standard deviation), which we account for via large-deviation bounds on per-group risk values. In particular, suppose *loss family* $\mathcal{F}\subseteq\mathcal{X}\to[0,r]$, take *per-group* distributions $\mathcal{D}_{1:g}$, and per-group samples $\boldsymbol{x}_1\sim\mathcal{D}_1^m,\ldots,\boldsymbol{x}_g\sim\mathcal{D}_g^m$. Now define per-group *supremum variances* and *Rademacher averages*[2]

$$\boldsymbol{\sigma}^2\doteq\left\langle\sup_{f\in\mathcal{F}}\mathop{\mathbb{V}}_{\mathcal{D}_1}[f],\ldots,\sup_{f\in\mathcal{F}}\mathop{\mathbb{V}}_{\mathcal{D}_g}[f]\right\rangle\ ,\quad\&\quad\mathfrak{R}\doteq\left\langle\mathfrak{R}_m(\mathcal{F},\mathcal{D}_1),\ldots,\mathfrak{R}_m(\mathcal{F},\mathcal{D}_g)\right\rangle\ ,$$

respectively. Then, for all $p\geq1$ and weights vectors $\boldsymbol{w}$, it holds that

$$\mathop{\mathbb{E}}_{\boldsymbol{x}}\Bigg[\underbrace{\sup_{f\in\mathcal{F}}\Big|\mathrm{M}_p\big(i\mapsto\mathop{\mathbb{E}}_{\mathcal{D}_i}[f];\boldsymbol{w}\big)-\mathrm{M}_p\big(i\mapsto\hat{\mathbb{E}}_{\boldsymbol{x}_i}[f];\boldsymbol{w}\big)\Big|}_{\textsc{Supremum Deviation of }\mathrm{M}_p(\cdot;\boldsymbol{w})}\Bigg]\leq2\mathrm{M}_p(\mathfrak{R};\boldsymbol{w})+\frac{r\ln(eg)}{3m}+\mathrm{M}_p(\boldsymbol{\sigma};\boldsymbol{w})\sqrt{\frac{2\ln(eg)}{m}}\ .\qquad(2)$$

[1] Indeed, the power mean comes from the same axiomatic foundations of cardinal welfare theory, the only difference being that to incentivize equitable redistribution, with welfare and allocations, we require $p\leq1$, whereas with harm and risk in machine learning, we require $p\geq1$. Thus while the specific choice of p parameterizes how we wish to trade off between helping the most harmed groups and minimizing overall harm, within these axioms, the power mean is the only reasonable objective for this task.

[2] The Rademacher average [1] is defined as $\hat{\mathfrak{R}}_m(\mathcal{F},\mathcal{D})\doteq\mathbb{E}_{\boldsymbol{x},\boldsymbol{\sigma}}\big[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^m\boldsymbol{\sigma}_i f(\boldsymbol{x}_i)\big]$, given $\boldsymbol{x}\sim\mathcal{D}^m$ and $\boldsymbol{\sigma}\sim\mathcal{U}^m(\pm1)$, and like its empirically centralized counterpart, it is used to control the supremum deviation.

In other words, the *supremum variances* and *Rademacher averages*, which are key quantities in controlling the supremum deviation of simple expectations, also control the supremum deviation of power means.

3. The Inter-Trace Variance [7]

Characterizing the fundamental difficulty of mean-estimation from dependent samples with a finite-sample variance statistic.

Let τ denote the *relaxation time*[3] of an ergodic Markov chain $\mathcal{M}$ over some space $\mathcal{X}$, and define the trace-length t inter-trace variance of a function $f : \mathcal{X} \to \mathbb{R}$ as

$$v^{(t)} \doteq \mathop{\mathbb{V}}_{\boldsymbol{x}_{1:n} \sim \mathcal{M}^{(t)}} \left[\frac{1}{\sqrt{t}} \sum_{i=1}^{t} f(\boldsymbol{x}_i) \right] \quad , \quad \& \quad v^{(\infty)} \doteq \lim_{t \to \infty} v^{(t)} \quad ,$$

where $\mathcal{M}^{(t)}$ denotes the stationary distribution of length-t traces drawn from $\mathcal{M}$. Then

$$\tfrac{1}{4} v^{(\infty)} \leq v^{(2\tau - 1)} \leq v^{(\infty)} \quad , \quad \text{and furthermore,} \quad \forall t \geq \tau : \; v^{(t)} \asymp v^{(\infty)} \quad . \tag{3}$$

The significance of these variance bounds is that we have *necessary* and *sufficient* bounds

$$m \in \mathbf{\Omega}\left(\frac{v^{(\tau)}}{\varepsilon^2} \log \frac{1}{\delta} \right) \quad , \quad \& \quad m \in \mathbf{O}\left(\left(\frac{\tau r}{\varepsilon} + \frac{v^{(\tau)}}{\varepsilon^2} \right) \log \frac{1}{\delta} \right) \quad ,$$

respectively, on the *sample complexity* i.e., the required number of Markov chain steps, of ε-δ additive mean estimation, via Markov chain central limit theorems [12] and Bernstein-type [11, 13] inequalities.

As a statistic on an infinite sample, the *asymptotic variance* $v^{(\infty)}$ is notoriously difficult to estimate or bound [10], even heuristically. On the opposite end of this spectrum, $v^{(1)}$, which is generally termed the *stationary variance* (as it reduces to the *variance* of the *stationary distribution* of $\mathcal{M}$), is easily estimated, but is insensitive to temporal dependence within the chain, and thus leads to suboptimal convergence guarantees. The *relaxation time* τ is the key ingredient here, as if it tells us that while the stationary variance $v^{(1)}$ may yield asymptotically suboptimal convergence rates, and $v^{(\infty)}$ may be prohibitively difficult to estimate, $v^{(\tau)}$ sits between them, matching both the simplicity of estimation of $v^{(1)}$ and the efficiency guarantees of $v^{(\infty)}$.

In Summary Inequality (1) is purely a statistical concept that beautifully and concisely summarizes decades of research in statistical learning theory in the i.i.d. setting. Inequality (2) bounds the sampling error of a key target for fairness-sensitive computing, relevant to machine learning, resource allocations, and algorithm selection, and concisely describes what should change when considering a population of *diverse subgroups*, rather than individuals or homogeneous populations. Finally, inequality (3) is again purely statistical, but now we have left the i.i.d. setting to provide guarantees for dependent (i.e., time-series) data under a Markovian assumption.

In my work, statistics are placed at the forefront, because it is my sincere belief (and, if I dare say so, my well-reasoned argument) that without considering the *statistics* of learning processes, copious additional avenues of failure arise. For this reason, I developed the *empirically centralized Rademacher average* to analyze and bound the generalization error of machine learning methods and other statistical processes, with both *asymptotically optimal* guarantees and *strong practical performance* in the small-sample setting, all from fully sample-dependent quantities with very few *a priori* assumptions. Similarly, no rigorous fairness guarantee can be complete without consideration of the various biases that may arise due to finite-sample estimation error. I thus argue for the *power mean* as a fair learning target, not only fundamentally from a welfare-economics perspective, but also due to its convenient statistical properties; here *generalization error* can be summarized by the gaps between the training set and distributional (or test set) power-mean

[3] The *relaxation time* τ is a standard measure of the degree of temporal autocorrelation between nearby samples drawn from the chain $\mathcal{M}$, and is defined as the *multiplicative inverse* of the *spectral gap* of $\mathcal{M}$, i.e., assuming $\mathcal{M}$ has second absolute eigenvalue λ_2, $\tau \doteq \frac{1}{1-\lambda_2}$.

objectives and per-group risk values, whereas methods based on *fairness constraints* [9] must consider the estimation error of both the *risk* or *accuracy* of a model, as well as the *demographic statistics* involved in the fairness constraints [14]. Finally, I investigated the trace variance $v^{(t)}$ partly to relax assumed *a priori* knowledge of $v^{(\infty)}$ in dependent data-analysis problems, but also because, as many have previously observed, the i.i.d. assumptions often made in theoretical machine learning are often unrealistic in practice, so in order to address the real problems of the world today, progress must be made in analyzing non-i.i.d. processes.

A Note on Statistical Estimation Each of the above three inequalities involve an *easily estimated* novel statistic or key quantity that is a variant of a well-known concept that is far more difficult to estimate from a finite sample. The above inequalities show that these novel variants can be used to control, bound, or estimate statistical error in much the same way as their classical counterparts. However, the real utility of these definitions only becomes apparent when one considers fully-empirical settings and attempts to relax *a priori* assumptions on such quantities, due to their vastly improved ease of estimation.

In particular, here the empirically centralized Rademacher average $\mathbb{E}_{\boldsymbol{x}}[\hat{\mathfrak{R}}_m(\mathcal{F} - \hat{\mathbb{E}}_{\boldsymbol{x}}[\mathcal{F}], \boldsymbol{x})]$ is a constant-factor approximation of the *centralized Rademacher average* $\mathbb{E}_{\boldsymbol{x}}\left[\hat{\mathfrak{R}}_m(\mathcal{F} - \mathbb{E}_{\mathcal{D}}[\mathcal{F}], \boldsymbol{x})\right]$, the *power mean welfare* $\mathrm{M}_p(\mathcal{S}; \boldsymbol{w})$ closely resembles the *additively separable welfare* $\mathrm{M}_p^p(\mathcal{S}; \boldsymbol{w})$, and the t-trace variance $v^{(t)}$ is a constant-factor approximation of the asymptotic variance $v^{(\infty)}$ for sufficiently large t. I show that the relevant quantities in each of the three inequalities can easily be estimated within ε-δ additive error, whereas their relatives are substantially harder to estimate. The statistical quantities, i.e., the Rademacher averages and myriad variance concepts, characterize sample complexity and sampling error, while being fully estimable from a sample, which leads to bounds on supremum deviations, see (1) & (2), as well as for mean estimation.

Concretely, I show *sub-gamma* type inequalities [3] for each quantity Z and an appropriate estimator or empirical quantity $\hat{Z}_m$. In particular, here Z describes the supremum deviation over $\mathcal{F}$, the supremum deviation over $p \geq 1$ power-means, or the mean of some f in a dependent sampling setting, and $\hat{Z}_m$ is the corresponding empirical supremum deviation or empirical mean. Assume $\hat{Z}_m$ is a statistic over g groups ($g = 1$ outside fairness settings), the relevant domain has range r, relaxation time τ ($\tau = 1$ in i.i.d. settings), $\delta \in (0, 1)$, and take $v^{(\tau)}$ to be the supremum τ-trace variance of any $f \in \mathcal{F}$ over any $\mathcal{D}_i$. It then holds that

$$\mathbb{P}\left(\left|Z - \hat{Z}_m\right| > \mathbf{O}\left(\frac{\tau r \log \frac{g}{\delta}}{m} + \sqrt{\frac{v^{(\tau)} \log \frac{g}{\delta}}{m}}\right)\right) \leq \delta \tag{4}$$

Inequality (4) is not a unified result under broad assumptions, but rather a summary of disparate results in [8, 5, 7] for the specific Z described above. The specifics of each case vary, and often sharper bounds are possible, but the key here is that each bound closely resembles the Bennett inequality [2], with logarithmic dependence on $\frac{1}{\delta}$, weak dependence on range, and strong (asymptotically dominant) variance dependence.

The eagle-eyed reader may observe that I have broken my central promise, and stated a fourth inequality, but I say it hardly counts, as it merely augments the central three, by specifying the rate at which the relevant quantities (i.e., the legion of Rademacher averages, variance concepts, risk values, expectations, and power means) may be estimated, and in any case, I could hardly conclude this piece without giving concrete idea of what these estimation guarantees look like after spilling so much ink on singing their praises!

Sample Complexity and Progessive Sampling For single-function mean estimation tasks, (4) may be applied verbatim, however the supremum deviation bounds require knowledge of Z (the expected supremum deviation). Fortunately, this is exactly the purpose of the Rademacher average, and in particular, (1) & (2) give such bounds. Furthermore, (empirically centralized) Rademacher averages, and even Monte-Carlo estimates thereof, wherein $\boldsymbol{\sigma}$ is also sampled, admit estimation guarantees quite similar to (4), thus these quantities can be fully-estimated from the sample. Dependence on the appropriate variance concept $v^{(\tau)}$ is also not a problem, as similar sampling bounds hold for the appropriate variance concepts, although since variance is a quadratic quantity, their forms are necessarily different. In both cases, in conjunction with a union bound, purely sample-dependent upper-bounds can be substituted into $\mathbb{E}[Z]$ or $v^{(\tau)}$ to remove the need for *a priori* variance assumptions, at no asymptotic cost.

These bounds on *estimation error* also immediately imply bounds on *sample complexity*, which characterize

the *sufficient sample size* to obtain an desired ε-δ approximation guarantee. In particular, we have that

$$\exists m \in \mathbf{O}\left(\left(\frac{\tau r}{\varepsilon} + \frac{v^{(\tau)}}{\varepsilon^2}\right)\log\frac{g}{\delta}\right) \text{ s.t. } \mathbb{P}\left(\left|Z - \hat{Z}_m\right| > \varepsilon\right) \leq \delta \ .$$

Of course, without *a priori* knowledge of $\mathbb{E}[Z]$ or $v^{(\tau)}$, this is not particularly useful. However, via the *progressive sampling* technique, we can relax even this requirement. In this method, we essentially "guess and check" with progressively doubling sample sizes (geometrically interpolating from the *minimum sufficient* to the *maximum necessary* sample sizes), at each iteration computing the appropriate bounds, and terminating when the desired estimation guarantee is met. Via this technique, with high probability, we can match the above sample complexity to within $\log\log$ factors[4], despite the algorithm being oblivious to the quantities that describe the appropriate statistical convergence rate (i.e., $\mathbb{E}[Z]$ and $v^{(\tau)}$).

A Taste of What is to Come My ongoing work is mostly focused on improving and broadening results in these three overarching directions. In particular, I am working on further improving small-sample guarantees with empirically centralized Rademacher averages, and developing new methods for bounding them, as well as converting the resulting *uniform bounds* to *relative bounds* for statistical tasks with both low and high frequency events. In the fair-learning setting, I am developing improved generalization bounds, tackling welfare and the $p < 1$ case,[5] as well as applications in *fair reinforcement learning* and other settings. There are also unanswered questions about the relationship between the computational complexity of fair learning ($g > 1$) and traditional learning ($g = 1$), and practical questions on how to efficiently optimize power-mean objectives. In dependent estimation settings, I am currently looking for problems for which trace variance estimation yields asymptotic improvement, in particular #P-hard counting problems and normalizing constant estimation for Gibbs distributions, and investigating learning settings to which these methods may be applied.

I am also highly interested in combining ideas from these seemingly-disparate areas of study. On some level, I have investigated most of the pieces of this puzzle somewhere in my work, but unfortunately, such a change in setting does not often admit analysis via simple composability. Preliminary study in this direction has encountered exciting technical challenges, for instance in adapting uniform convergence bounds or analysis of empirical power means to non-i.i.d. settings, and I am excited for what the future holds!

References

[1] Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

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

[3] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[4] Cyrus Cousins. Novel concentration of measure bounds with applications to fairness in machine learning. *Brown University*, 2020.

[5] Cyrus Cousins. An axiomatic theory of provably-fair welfare-centric machine learning. *arXiv preprint arXiv:2104.14504*, 2021.

[6] Cyrus Cousins. *Bounds and Applications of Concentration of Measure in Fair Machine Learning and Data Science*. PhD thesis, Brown University, 2021.

[7] Cyrus Cousins, Shahrzad Haddadan, and Eli Upfal. Making mean-estimation more efficient using an MCMC trace variance approach: DynaMITE. *arXiv preprint arXiv:2011.11129*, 2020.

[8] Cyrus Cousins and Matteo Riondato. Sharp uniform convergence bounds through empirical centralization. In *Advances in Neural Information Processing Systems*, 2020.

[4] The doubling granularity introduces only constant-factor cost; the $\log\log$ factor rather comes from the *union bound* over each sample size considered; in particular, this cost is $\ln\lceil\log_2 \frac{m^{\uparrow}}{m^{\downarrow}}\rceil$, where $m^{\downarrow}$ and $m^{\uparrow}$ are *a priori* bounds on the *minimum sufficient* and *maximum necessary* sample sizes, and the ratio between them is usually polynomial in ε^{-1} and all other variables.

[5] Though it seems a rather uninspired challenge, power-means with $p < 1$ differ substantially from those with $p > 1$, and are significant for learning fair allocations. Furthermore, this case is technically challenging, primarily due to non-Lipschitz behavior for $p \in [0, 1)$, and thus requires more sophisticated analysis tools, subtler conditions, and/or relaxed estimation guarantees.

[9] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

[10] Charles J Geyer. Practical Markov chain Monte Carlo. *Statistical science*, pages 473–483, 1992.

[11] Bai Jiang, Qiang Sun, and Jianqing Fan. Bernstein's inequality for general Markov chains. *arXiv preprint arXiv:1805.10721*, 2018.

[12] Claude Kipnis and SR Srinivasa Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19, 1986.

[13] Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electron. J. Probab.*, 20, 2015.

[14] Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.