Wireless Networks and the Utopia of Peak Performance

Martin Haenggi

University of Notre Dame, IN, USA
ETH Zürich, Switzerland

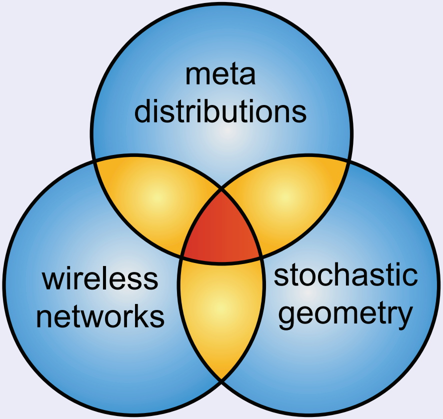
ISI, ETHZ

Feb. 7, 2022

Available at http://www.nd.edu/~mhaenggi/talks/ethz22.pdf

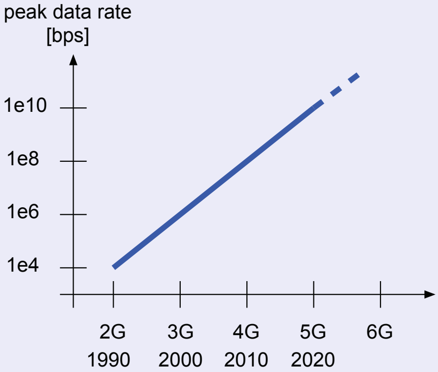
Overview

The three themes

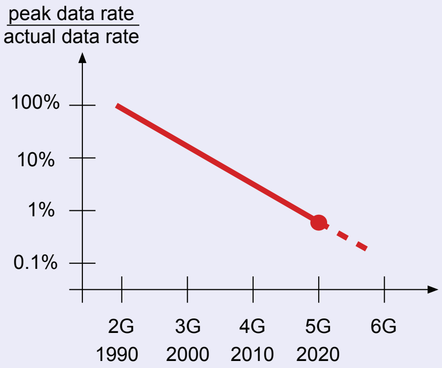


Wireless network evolution

The usual plot: Peak rates in the cellular generations

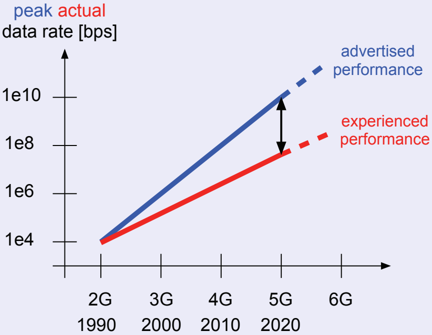


Our focus: Peak vs. actual rates



Actual data rate: Performance of the 5-percentile user (maximum rate of the bottom 5% of users), called "user experienced data rate" in 5G parlance.

Advertised (peak) and experienced (5%) rates



There is increasing uncertainty about the performance, and it is near-impossible to achieve peak rates.
How can we analyze percentile performance and improve it?

And encourage operators to advertise and publish experienced performance?

The law of low expectations

Analogies

- Surprisingly, users tend to accept data rate promises of "up to x Gbps" by cellular devices and operators.
- Compare with a job advertisement: "Take this a wonderful job—peak pay at our company is $10M/year."
- Or with an ad for an electric car: "This vehicle has a peak range of 10 Mm per charge."
- Only in one other industry "up to" statements are generally accepted: the lotteries.

Interference

The SINR and SIR

Interference—the cumulative power of undesired signals—is the key performance-limiting factor in many modern wireless systems.
The quantity that determines the throughput (rate), delay, and reliability of a link is the signal-to-interference-plus-noise ratio

$$\mathrm{SINR} = \frac{S}{I + W}.$$

If $I \gg W$, we can ignore the noise and focus on just the SIR.

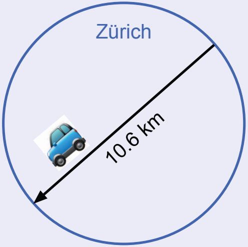
Key issue: Peak rate calculations assume no interference.

Consequence

But if interference limits the performance, users can only achieve rates much smaller than peak rates.

One more analogy

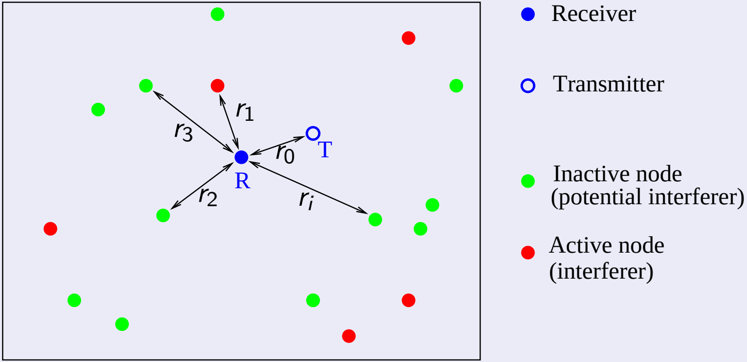
"Driving in Zürich is very efficient—it takes only 12 min to traverse the entire city!"



Indeed, with a speed limit of 50 km/h, it takes $\frac{10.6\text{ km}}{53\text{ km/h}} = 12$ min.

Such calculation ignores interference from other cars, traffic lights, stop signs, pedestrians, trams, construction, and Covid demonstrations.

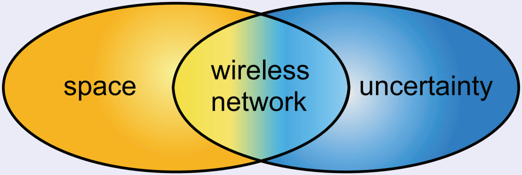
Generic wireless network



Transceivers are embedded in $\mathbb{R}^2$ or $\mathbb{R}^3$.
Interference is a function of distances and propagation conditions, most of which are unknown at the transceivers.

Wireless differentiae

"Wireless" lies at the intersection of space and uncertainty

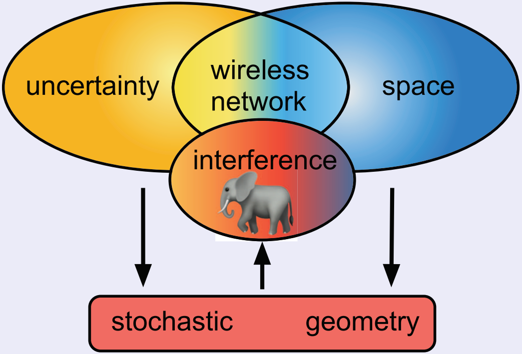


Concurrent transmissions on the same frequency need to be separated in space.

Space is the critical resource, and wireless networking is about space and uncertainty management.

E.g.: Base station placement, user association, power control, small cells, feedback, (massive) MIMO, CoMP, SIC, CSMA, NOMA, IRS, diversity schemes, edge computing, caching.

Analytical tool: Stochastic geometry



Stochastic geometry provides both the models and the analytical tools for wireless networks.
In particular, it is the only theory that permits a characterization of the interference.

Randomness in wireless networks

What do we model as random:
(a) What is very hard to compute: multi-path propagation (small-scale fading)
(b) What cannot possibly be known at the transmitter or receiver: distances and channels to interfering transmitters

Standard model for received signal strength S over a wireless link:

$$S = Phr^{-\alpha}$$

P is the transmit power, h the channel state (fading), r the link distance, and α the path loss exponent.
To model the transceiver locations, we use a point process $\Phi = \{x_1, x_2, \ldots\} \subset \mathbb{R}^2$. The SIR then has an expression of the form

$$\mathrm{SIR} = \frac{P_0 h_0 \|x_0\|^{-\alpha}}{\sum_{x \in \Phi \setminus \{x_0\}} P_x h_x \|x\|^{-\alpha}}.$$

Parsing the SIR

$$\mathrm{SIR} = \frac{P_0 h_0 \|x_0\|^{-\alpha}}{\sum_{x\in\Phi\setminus\{x_0\}} P_x h_x \|x\|^{-\alpha}}.$$

Random elements: h_0, h_x (iid, time-varying), Φ (spatial, static). The receiver (and transmitter) can learn $h_0\|x_0\|^{-\alpha}$ but not Φ or h_x.

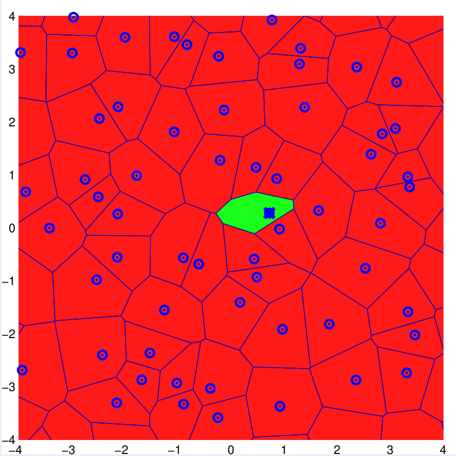
Point process Φ: Central object in stochastic geometry. It can be viewed as a random set or a random counting measure, where $\Phi(B) \in \mathbb{N}_0$ is the number of points in the Borel $B \subset \mathbb{R}^2$.
Prominent example: Poisson point process, where $\Phi(B)$ is Poisson with mean $\lambda|B|$ and $\Phi(B_1)$ and $\Phi(B_2)$ are independent if $B_1 \cap B_2 = \emptyset$.

Key advantages of spatial modeling (cf. fading models):

- Tractability: A large (infinite) number of deterministic locations would result in very unwieldy expressions.
- Generality: By averaging over Φ, results are obtained that capture all the likely network configurations.

Example: Poisson cellular networks

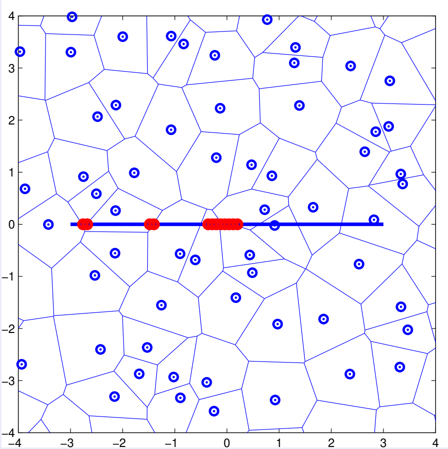
Fully loaded network (frequency reuse 1)

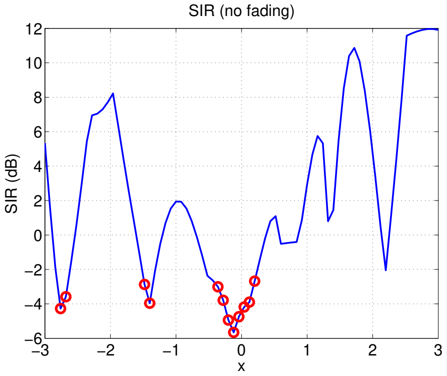


All base stations form a Poisson point process (PPP) and use the entire bandwidth. The serving base station is the closest one.

$$\mathrm{SIR} = \frac{S}{I}$$

A walk through the single-tier cellular network





In the red locations, performance is poor (low SIR).

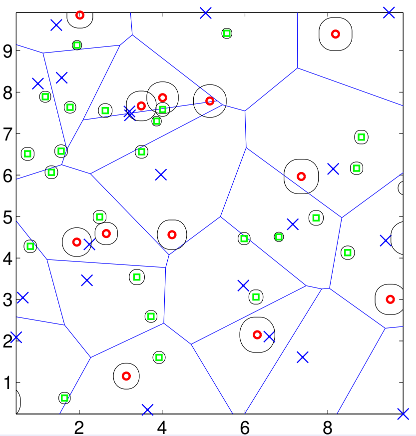
SIR fluctuation including fading

Due to ergodicity, the fraction of the curve that is above the threshold θ corresponds to the ccdf of the SIR at θ:

$$\bar{F}_{\mathsf{SIR}}(\theta) \triangleq \mathbb{P}(\mathsf{SIR} > \theta)$$

The SIR ccdf is often equated with the success probability (reliability) of a transmission, where θ captures the rate.

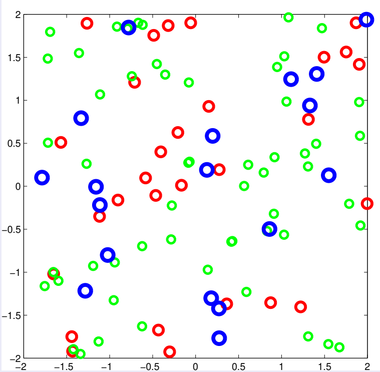
A multi-tier model



Heterogeneous network with macro-base stations, micro-base stations and femtocells. Users connect to the one with the strongest received power (on average).

The HIP model

The HIP (homogeneous independent Poisson) model



Base stations in each tier form independent PPPs of densities λ_i and transmit powers P_i.
Here $\lambda_i = 1, 2, 3$ for the blue, red, and green tiers.

The SIR is

$$\mathrm{SIR} = \frac{P_0 h_0 \|x_0\|^{-\alpha}}{\sum_{x \in \Phi \setminus \{x_0\}} P_x h_x \|x\|^{-\alpha}}$$

SIR distribution for HIP downlink

Consider a user at the origin o. This user, when averaging over the BS location point process, becomes the typical user. Assume Rayleigh fading, *i.e.*, h_0, h_x are iid exponential with mean 1.

Remarkably, the SIR distribution at the typical user is independent of the number of tiers, their densities, and their power levels

$$\bar{F}_{\mathrm{SIR}}(\theta) = \frac{1}{{}_2F_1(1, -\delta; 1-\delta; -\theta)}, \quad \delta \triangleq 2/\alpha.$$

$$\text{For } \delta = 1/2 \ (\alpha = 4): \quad \bar{F}_{\mathrm{SIR}}(\theta) = \frac{1}{1 + \sqrt{\theta}\arctan\sqrt{\theta}}$$

For other point processes, this ccdf can be tightly approximated as $\bar{F}_{\mathrm{SIR}}(\theta/G)$ for $G \in (0, 2.2]$. For a square lattice, $G \approx 2$.

Interpretation of the SIR ccdf

The SIR distribution $\bar{F}_{\mathrm{SIR}}$ can be interpreted as:

- The ensemble average (over fading and point process Φ) $\mathbb{E}\mathbf{1}(\mathrm{SIR}(x) > \theta)$ at any location x.
- The spatial average of the area fraction in which $\mathrm{SIR}(x) > \theta$ in any realization of Φ:

$$\lim_{r\to\infty} \frac{1}{\pi r^2} \int_{b(o,r)} \mathbf{1}(\mathrm{SIR}(x) > \theta)\mathrm{d}x$$

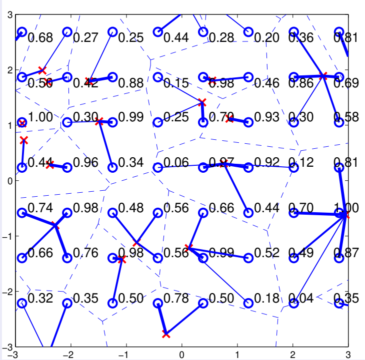
  The locations for which $\mathrm{SIR}(x) > \theta$ are changing quickly over time.

Limitation

- The SIR ccdf does not reveal any information about the disparity between the users and, by consequence, about user percentiles.
- Despite its appearance as a distribution, it is just an average that muddles temporal and spatial randomness.

Per-user SIR distribution in cellular network

For each user u, calculate $P_{\rm s}^{(u)} = \mathbb{P}(\mathrm{SIR}_u > \theta \mid \Phi) = \mathbb{E}_h \mathbf{1}(\mathrm{SIR}_u > \theta)$:

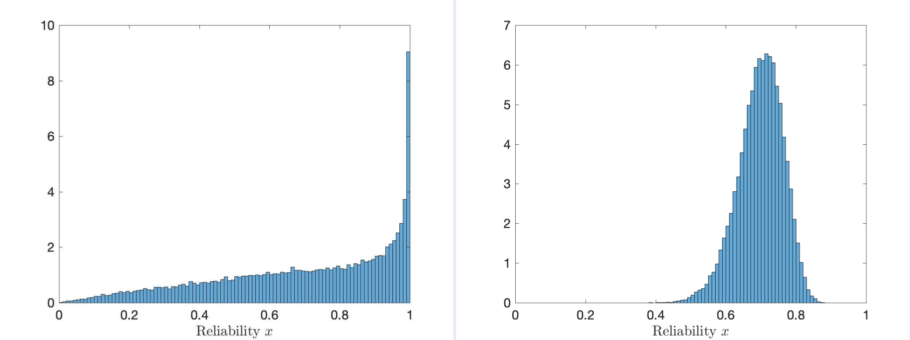


Per-user reliabilities, averaged only over the fading, for $\alpha = 4$, $\theta = 1$.

The SIR distribution at $\theta = 1$ is the average $\bar{F}_{\mathrm{SIR}}(1) = \mathbb{E}P_{\rm s}^{(o)} = 0.56$.
But to assess the disparity between the users and the 5-percentile performance, we need the distribution of $P_{\rm s}^{(o)}$.

The flaw of averages

Reliability histograms

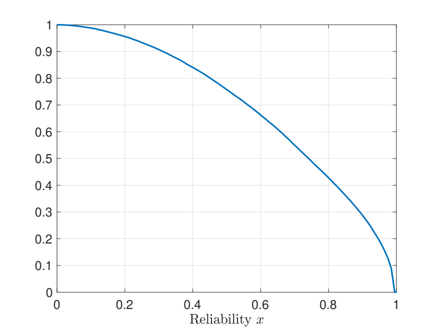


Case 1 | Case 2

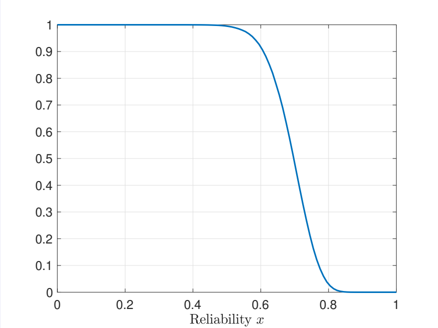
In both cases, the average success probability is 0.70. Clearly, a distribution as in Case 2 is much more desirable, but the standard analysis makes no difference between the two.

Per-user performance (cumulative distribution)

What fraction of users achieve a reliability of x at a given rate?



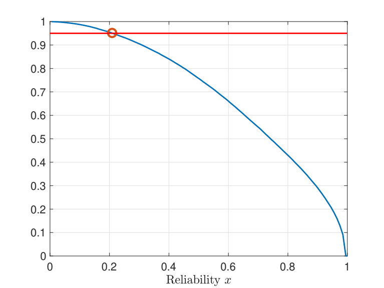
Case 1



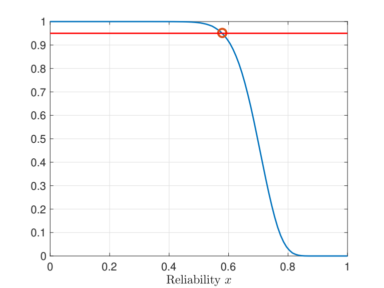
Case 2

Conversely, what is the reliability of the 5-percentile user in both cases?

5-percentile reliability



Case 1: x(0.95)=0.21

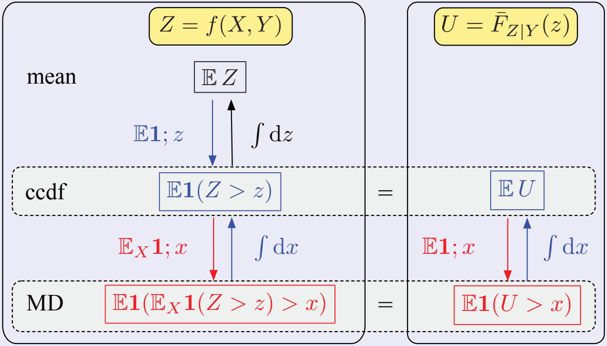


Case 2: x(0.95)=0.58

Techniques that boost the mean may be harmful to the 5% user.
Conversely, others may have little effect on the mean but increase the fairness (concentration).

The concept of meta distributions

Let $f : \mathbb{R}^2 \mapsto \mathbb{R}$ and X, Y be random variables.



A meta distribution (MD) is the distribution of a conditional distribution. It reveals how X, Y individually affect Z. Written differently,

$$\bar{F}_{[\![Z|Y]\!]}(z,x) = \mathbb{P}(\mathbb{P}(Z > z \mid Y) > x) = \mathbb{P}(\mathbb{P}_X(Z > z) > x).$$

Toy example

MD of ratio of exponential random variables

Let X, Y be independent exponential with mean 1 and $1/\mu$, respectively. The ccdf of $Z \triangleq X/Y$ is $\bar{F}_Z(z) = \frac{\mu}{z+\mu}$ and $\mathbb{E}Z$ does not exist.
$U \triangleq \bar{F}_{Z|Y}(z) = \mathbb{E}\mathbf{1}(Z > z \mid Y) = e^{-Yz}$ and

$$\bar{F}_{[\![Z|Y]\!]}(z, x) = \mathbb{P}(U > x) = \mathbb{P}(e^{-Yz} > x) = 1 - x^{\mu/z}.$$

The ccdf of Z is retrieved by

$$\mathbb{E}U = \int_0^1 (1 - x^{\mu/z})\mathrm{d}x = \frac{\mu}{z+\mu}.$$

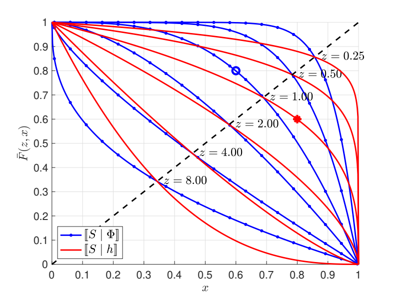
The dual MD is the inverse (in x)

$$\bar{F}_{[\![Z|X]\!]}(z, x) = (1 - x)^{z/\mu}.$$

Toy application

Signal power in uplink Poisson cellular network

Base stations form a PPP Φ of intensity λ, and users are served by the nearest one whose distance R is Rayleigh with mean $1/(2\sqrt{\lambda})$.
The signal power received is $S = h/R^2$, where h is exponential with mean 1.



$$\bar{F}_{[\![S|\Phi]\!]}(z,x) = \bar{F}_{[\![Z|Y]\!]}(z,x) = 1 - x^{\lambda\pi/z}$$

This is the fraction of users for which $S_u > z$ with probability at least x, for each realization of Φ.

80% of the users achieve $S = 1$ with probability at least 0.6.

The MD achieves a clean separation of temporal and spatial randomness.

SIR meta distribution in downlink Poisson cellular networks

For the same model but in the downlink, if we only consider the nearest interfering base station, with path loss $\ell(r) = r^{-2/\delta}$,

$$\bar{F}_{[\![\mathrm{SIR}'|\Phi]\!]}(\theta, x) = \min\left\{1, \left(\frac{1-x}{\theta x}\right)^{\delta}\right\}.$$

If all interferers are considered, a closed-form solution does not exist. We can, however, calculate the moments of the conditional distribution. For $U \triangleq \bar{F}_{\mathrm{SIR}|\Phi}(\theta) = \mathbb{P}(\mathrm{SIR} > \theta \mid \Phi)$,

$$M_b \triangleq \mathbb{E}(U^b) = \frac{1}{{}_2F_1(b, -\delta; 1-\delta; -\theta)}, \quad b \in \mathbb{C}.$$

The prior result for $\bar{F}_{\mathrm{SIR}}$ is just the special case $b = 1$.

Can we retrieve the distribution from the moments?

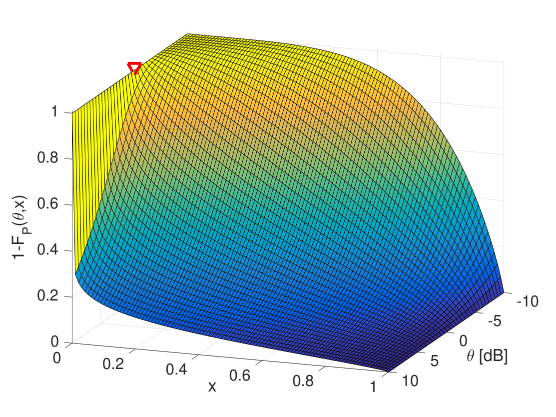
Mapping moments to distributions

- Equipped with imaginary moments, we have (Gil-Pelaez),

$$\bar{F}_{[\![\mathrm{SIR}|\Phi]\!]}(\theta, x) = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \frac{\Im(e^{-jt \log x} M_{jt})}{t} \mathrm{d}t.$$

  This is tricky to evaluate. There are no closed-form solutions.
- If the moments $b \in \mathbb{N}$ are known, it is called the Hausdorff moment problem. If only a finite number of moments are known, it is truncated and has infinitely many solutions (in most cases).
- Infima and suprema can be calculated with some effort, but calculating good approximations efficiently is ongoing research.
- The beta distribution (with matching M_1 and M_2) often gives good results.
- The variance $M_2 - M_1^2$ gives basic information on how concentrated the performance is.

Example (PPP, Rayleigh fading, $\alpha = 4$)

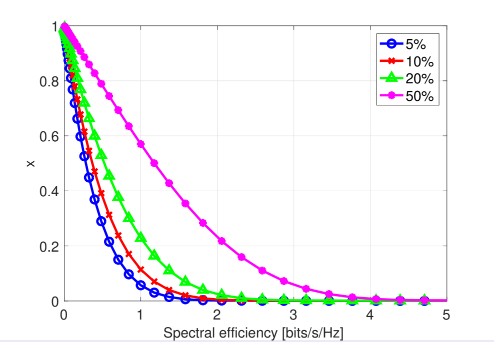


$$\bar{F}_{[\![\mathrm{SIR}|\Phi]\!]}(\theta, x) = \mathbb{P}(\mathbb{P}(\mathrm{SIR} > \theta \mid \Phi) > x)$$

$=$ Fraction of users who achieve an SIR of θ with probability at least x.

Trading off rate and reliability for user percentiles

Contours in the MD plot give the pairs (θ, x) that a given fraction of users achieve. With the (optimistic) mapping $\log_2(1+\theta)$ of θ to the spectral efficiency, we obtain the rate-reliability tradoff.



The significant gap between 5% and 50% user performance should be narrowed.

Insights

The meta distribution analysis reveals:

- Densification has little effect on the SIR (but increases the capacity).
- Power control in the cellular uplink barely affects the SIR distribution (average) but it drastically improves the 5-percentile performance.
- Base station cooperation (CoMP) where all users combine signals from multiple base stations improves the average but widens the gap to the 5-percentile user. Instead, CoMP should only be used for edge users (location-specific CoMP).
- Non-orthogonal multiple access (NOMA) should be restricted to users close to base stations.

Ergodic spectral efficiency

Motivation

- The reliability-based framework is useful for short(er) messages and low-latency situations.
- For longer messages (codewords) transmitted over larger bandwidths or many antennas or using hybrid ARQ, an ergodic point of view is warranted, where averaging over fading occurs and the rate is chosen such that no outages occur.
- If the resulting rate is normalized by the bandwidth, it is called ergodic spectral efficiency, measured in bps/Hz.

First attempt

Inspired by Shannon, calculate

$$\mathbb{E}\log_2(1+\mathrm{SIR}) = \mathbb{E}\log_2\left(1+\frac{h_0\|x_0\|^{-\alpha}}{\sum_{x\in\Phi\setminus\{x_0\}} h_x\|x\|^{-\alpha}}\right).$$

- This is (again) just an average that lumps together all sources of randomness.
- There is an underlying assumption of Gaussian signaling (reasonable).
- More importantly, it assumes that the receiver knows all the fading coefficients h_x. This results in a (fairly loose) upper bound.

Better approach

Condition on the point process first and eliminate the fading in the SIR expression.

Ergodic spectral efficiency

Let
$$\rho \triangleq \frac{\|x_0\|^{-\alpha}}{\sum_{x\in\Phi\setminus\{x_0\}}\|x\|^{-\alpha}}$$
be the SIR (of the user at the origin) without fading. It captures the network geometry. Next, let

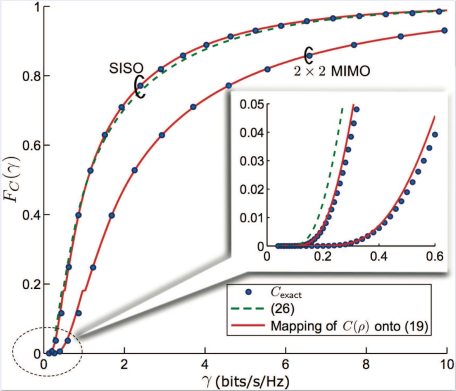
$$C(\rho) \triangleq \mathbb{E}_h(\log_2(1+h\rho))$$

be the ergodic spectral efficiency given the point process. For Rayleigh fading, $C(\rho)=e^{1/\rho}E_1(1/\rho)$, where E_1 is the exponential integral.

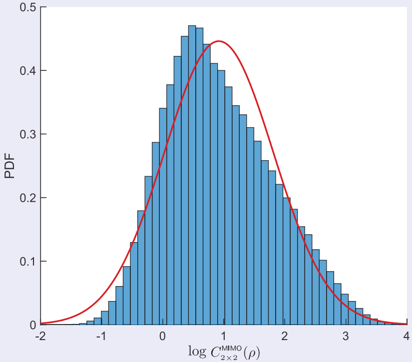
Ignoring the fading of the interferers yields a tight lower bound.

Now we can use stochastic geometry to find the ccdf $\bar{F}_{C(\rho)}(\gamma)=\mathbb{P}(C(\rho)\leq\gamma)$. This is a degenerate meta distribution.

Distribution of the ergodic spectral efficiency



With SISO, essentially no user gets less than 0.18 bps/Hz. With 2x2 MIMO, no user gets less than 0.3 bps/Hz.



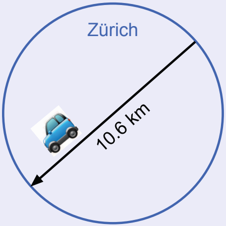
Interesting observation: Spectral efficiencies are essentially lognormal.

Proper averaging

Averaging over time vs. averaging over users

- At 3am, it may take us only 24 min to traverse the city.
- At 5pm, it probably takes us 48 min.
- The average is 47.5 min.

Reason: There are 50× more cars on the street at 5pm than at 3am.



Similarly, since many more users are active during daytime, nighttime performance is essentially irrelevant. So we need to focus on fully loaded conditions.

Same for space—crowds have more weight than isolated users.

Conclusions

- Advertised (peak) rates are utopian. The actual performance is a vanishing fraction.
- Sensibly evaluated averages are better but do not reveal what most users can expect. Instead, the focus should be on percentile performance.
- Space is the critical resource to manage interference, and stochastic geometry is the natural analytical tool. Combined with the meta distribution concept, it allows the characterization of the complete distribution of the performance metrics.
- By focusing on the 5-percentile performance, we can engineer wireless networks to behave more deterministically.
  And hold operators more accountable.

The last word

"At the pullout, you get 5% of the peak rate."

Slides available at: www.nd.edu/~mhaenggi/talks/ethz22.pdf