Propagation and Provenance of Uncertainty in Cyberinfrastructure-Related Data Processing

Vladik Kreinovich

Cyber-ShARE Center
University of Texas at El Paso
El Paso, TX 79968, USA
contact email vladik@utep.edu

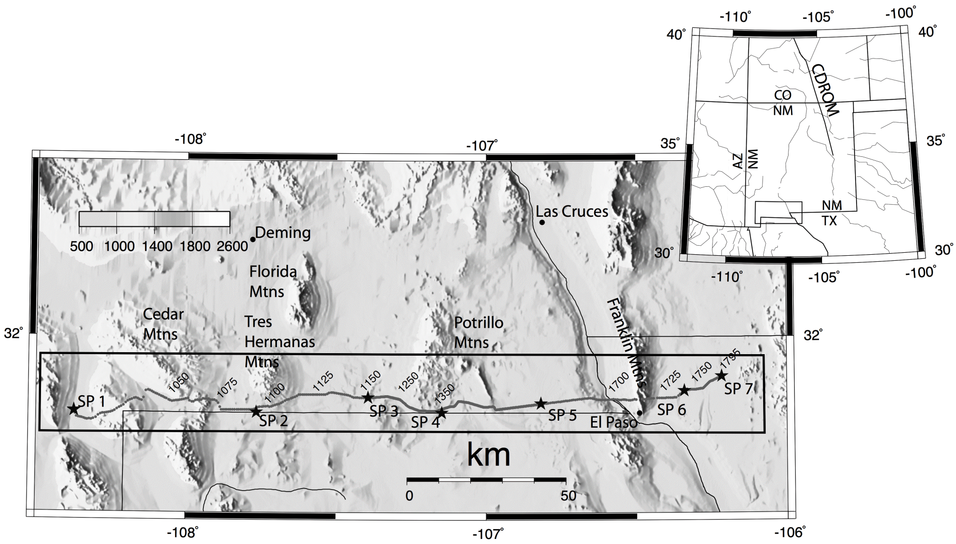
1. Outline

- Need for uncertainty estimation in cyberinfrastructure-related data processing.
- Geophysical case study: need to go beyond traditional techniques.
- Estimating uncertainty and spatial resolution.
- Combining different types of uncertainty: model fusion, with additional continuous vs. discrete problem.
- This is all based on known measurement results, how can be better plan the measurements?
- Optimal location of a sensor, on the example of a meteorological tower.
- Optimal placement of stationary sensors.
- Optimal trajectories of mobile sensors, on the example of UAV-based sensors.

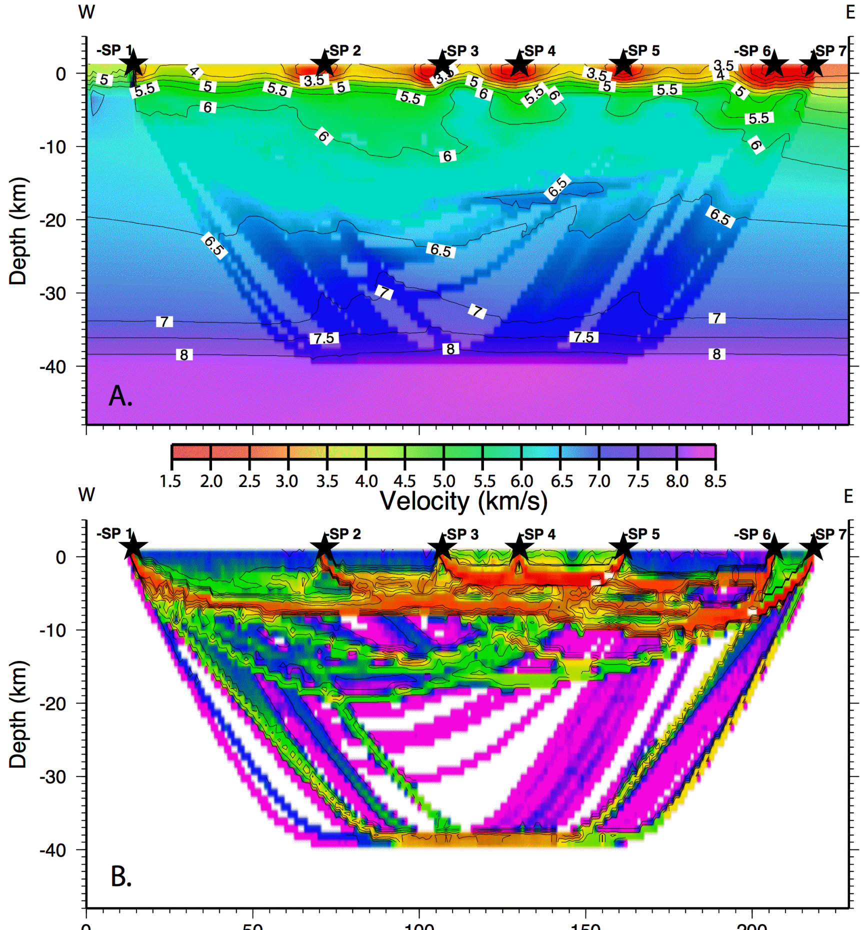
2. Need for Uncertainty Estimation in Cyberinfrastructure-Related Data Processing

- *In the past:* communications were much slower.
- *Conclusion:* use centralization.
- *At present:* communications are much faster.
- *Conclusion:* use cyberinfrastructure.
- *Related problems:*
  - gauge the the uncertainty of the results obtained by using cyberinfrastructure;
  - which data points contributed most to uncertainty;
  - how an improved accuracy of these data points will improve the accuracy of the result.
- *We need:* algorithms for solving these problems.

3. Case Study: Seismic Inverse Problem in the Geosciences

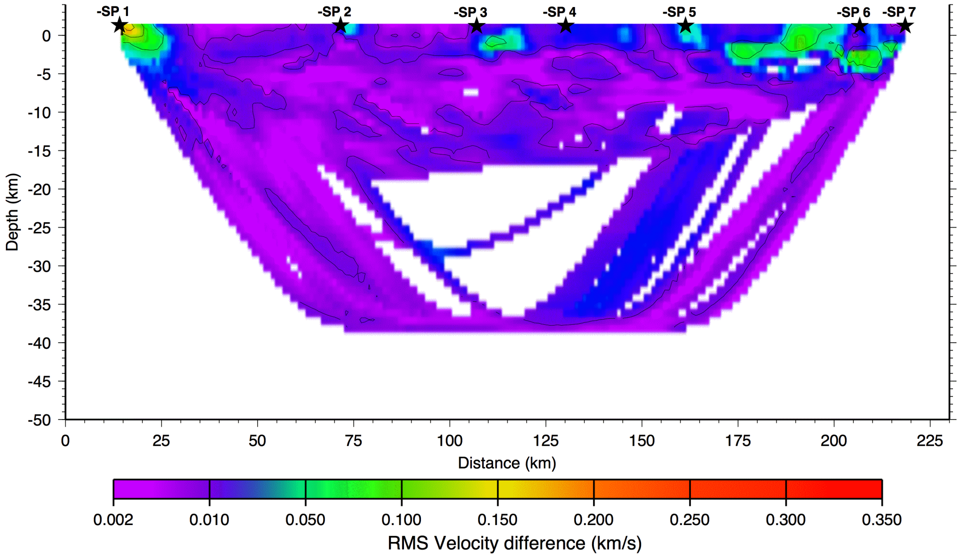


Outline
Need for Uncertainty . . .
Case Study: Seismic . . .
Need to go Beyond . . .
Model Fusion: We . . .
Optimal Stationary . . .
Dynamic Sensors: . . .
Home Page
Title Page
Page 4 of 24
Go Back
Full Screen
Close
Quit



Outline
Need for Uncertainty . . .
Case Study: Seismic . . .
Need to go Beyond . . .
Model Fusion: We . . .
Optimal Stationary . . .
Dynamic Sensors: . . .
Home Page
Title Page
Go Back
Full Screen
Close
Quit

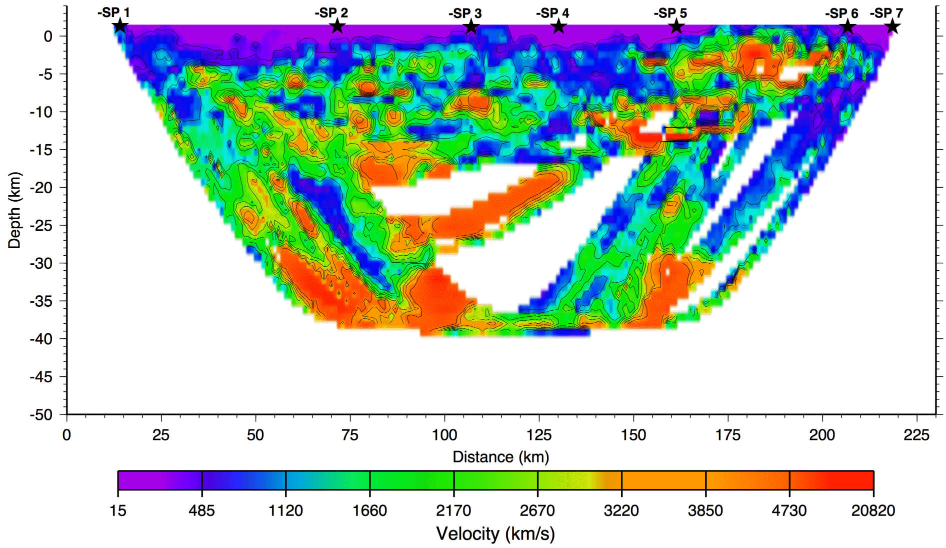
4. Need to go Beyond Traditional Probabilistic Techniques



5. Towards Interval Approach

- Manufacturer of the measuring instrument (MI) supplies Δ_i s.t. $|\Delta x_i| \le \Delta_i$, where $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$.
- The actual (unknown) value x_i of the measured quantity is in the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.
- *Probabilistic uncertainty:* often, we know the probabilities of different values $\Delta x_i \in [-\Delta_i, \Delta_i]$.
- *How probabilities are determined:* by comparing our MI with a much more accurate (standard) MI.
- *Interval uncertainty:* in two cases, we do not determine the probabilities:
  - cutting-edge measurements;
  - measurements on the shop floor.
- In both cases, we only know that $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

6. Estimating Uncertainty, Second Try: Interval Approach



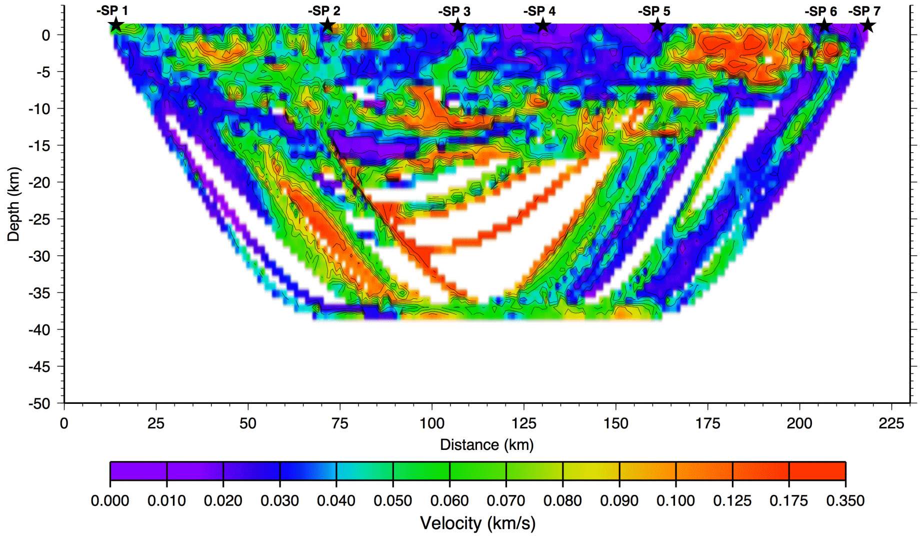
7. Towards a Better Estimate

- *Linearization:* $\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i$, where $c_i \stackrel{\text{def}}{=} \dfrac{\partial f}{\partial x_i}$.
- *Formulas:* $\sigma^2 = \sum_{i=1}^{n} c_i^2 \cdot \sigma_i^2$, $\Delta = \sum_{i=1}^{n} |c_i| \cdot \Delta_i$.
- *Numerical differentiation:* n iterations, too long.
- *Monte-Carlo approach:* if Δx_i are Gaussian w/σ_i, then $\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i$ is also Gaussian, w/desired σ.
- *Advantage:* # of iterations does not grow with n.
- *Interval estimates:* if Δx_i are Cauchy, w/$\rho_i(x) = \dfrac{\Delta_i}{\Delta_i^2 + x^2}$, then $\Delta y = \sum_{i=1}^{n} c_i \cdot \Delta x_i$ is also Cauchy, w/desired Δ.

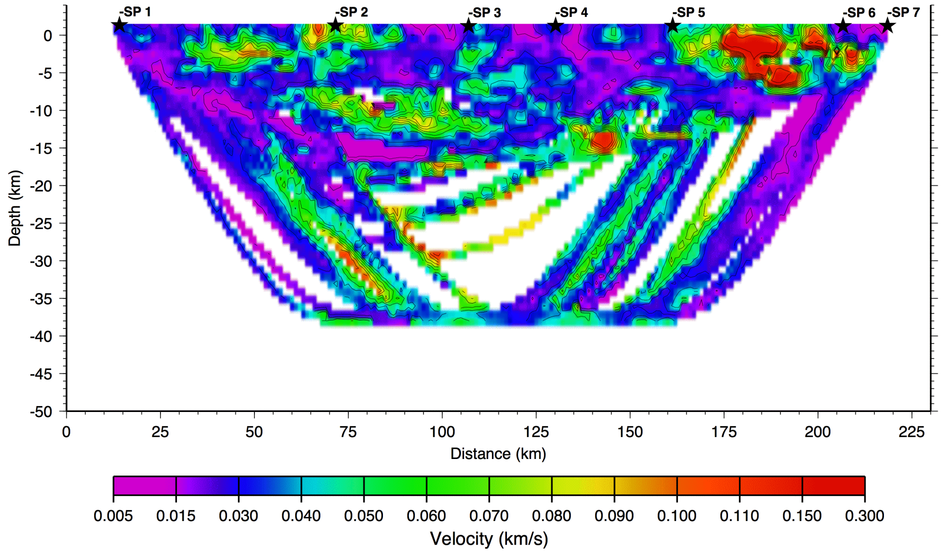
8. A New (Heuristic) Approach

- *Problem:* guaranteed (interval) bounds are too high.
- *Gaussian case:* we only have bounds guaranteed with confidence, say, 90%.
- *How:* cut top 5% and low 5% off a normal distribution.
- *New idea:* to get similarly estimates for intervals, we "cut off" top 5% and low 5% of Cauchy distribution.
- *How:*
  - find the threshold value x_0 for which the probability of exceeding this value is, say, 5%;
  - replace values x for which $x > x_0$ with x_0;
  - replace values x for which $x < -x_0$ with $-x_0$;
  - use this "cut-off" Cauchy in error estimation.
- *Example:* for 95% confidence level, we need $x_0 = 12.706$.

9. Heuristic Approach: Results with 95% Confidence Level



10. Heuristic Approach: Results with 90% Confidence Level



11. Model Fusion: We Also Have Different Spatial Resolution

- In many situations, different models have not only different accuracy, but also different spatial resolution.
- *Example:*
  - seismic data leads to higher spatial resolution estimates of the density at different locations, while
  - gravity data leads to lower-spatial resolution estimates of the same densities.
- *Towards precise formulation of the problem:*
  - High spatial resolution estimates correspond to small spatial cells.
  - A low spatial resolution estimate is affected by several neighboring spatial cells.

12. Estimates of High and Low Spatial Resolution: Illustration

$\widetilde{x}_1 = 2.0$	$\widetilde{x}_2 = 3.0$
$\widetilde{x}_3 = 5.0$	$\widetilde{x}_4 = 6.0$

$$\widetilde{X}_1 = 3.7$$

13. Numerical Example: Discussion

- We assume that the low spatial resolution estimate is accurate ($\sigma_l \approx 0$).
- So, the average of the four cell values is equal to the result $\widetilde{X}_1 = 3.7$ of this estimate:

$$\frac{x_1 + x_2 + x_3 + x_4}{4} \approx 3.7.$$

- For the high spatial resolution estimates $\widetilde{x}_i$, the average is slightly different:

$$\frac{\widetilde{x}_1 + \widetilde{x}_2 + \widetilde{x}_3 + \widetilde{x}_4}{4} = \frac{2.0 + 3.0 + 5.0 + 6.0}{4} = 4.0 \neq 3.7.$$

- *Reason:* high spatial resolution estimates are much less accurate: $\sigma_h = 0.5$.
- We use the low spatial resolution estimate to "correct" the high spatial resolution estimate.

14. The Result of Model Fusion

$\widetilde{x}_1 \approx 1.89$	$\widetilde{x}_2 \approx 2.79$
$\widetilde{x}_3 \approx 4.62$	$\widetilde{x}_4 \approx 5.53$

- The arithmetic average of these four values is equal to
$$\frac{x_1 + x_2 + x_3 + x_4}{4} \approx \frac{1.89 + 2.79 + 4.62 + 5.53}{4} \approx 3.71.$$
- So, within our computation accuracy, it coincides with the low spatial resolution estimate $\widetilde{X}_1 = 3.7$.

15. Optimal Stationary Sensor Placement: Case Study

- *Objective:* select the best location of a sophisticated multi-sensor meteorological tower.
- *Constraints:* we have several criteria to satisfy.
- *Example:* the station should not be located too close to a road.
- *Motivation:* the gas flux generated by the cars do not influence our measurements of atmospheric fluxes.
- *Formalization:* the distance x_1 to the road should be larger than a threshold t_1: $x_1 > t_1$, or $y_1 \stackrel{\text{def}}{=} x_1 - t_1 > 0$.
- *Example:* the inclination x_2 at the tower's location should be smaller than a threshold t_2: $x_2 < t_2$.
- *Motivation:* otherwise, the flux determined by this inclination and not by atmospheric processes.

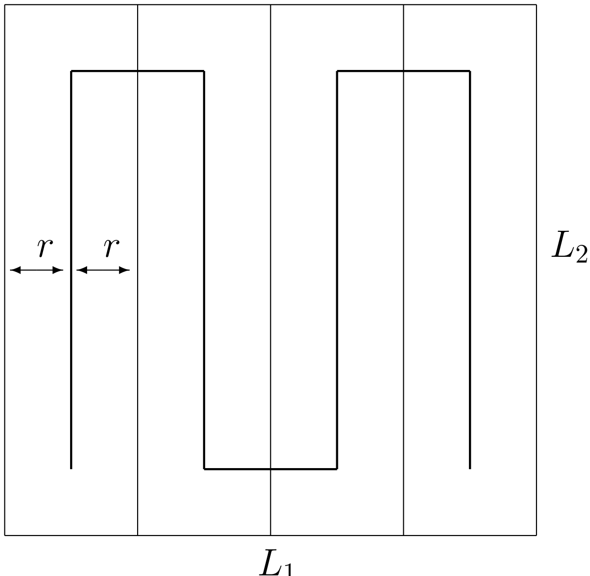
16. Main Result

- *Case study:* meteorological tower.
- *This case* is an example of multi-criteria optimization, when we need to maximize several objectives $x_1, \ldots, x_n$.
- *Traditional approach* to multi-objective optimization: maximize a weighted combination $\sum\limits_{i=1}^{n} w_i \cdot x_i$.
- *Specifics of our case:* constraints $x_i > x_i^{(0)}$ or $x_i < x_i^{(0)}$.
- *Equiv.:* $y_i > 0$, where $y_i \stackrel{\text{def}}{=} x_i - x_i^{(0)}$ or $y_i = x_i^{(0)} - x_i$.
- *Limitations* of using the traditional approach under constraints.
- *Scale invariance:* a better description.
- *Main result:* scale invariance leads to a new approach: maximize $\sum\limits_{i=1}^{n} w_i \cdot \ln(y_i) = \sum\limits_{i=1}^{n} w_i \cdot \ln \left|x_i - x_i^{(0)}\right|$.

17. Dynamic Sensors: Need for an Optimal Trajectory

- *Task:* cover all the points points from a given area.
- *Problem:* UAVs have limited flight time.
- *Consequence:* minimize the flight time among all covering trajectories.
- *Geometric reformulation:* we need a trajectories with the smallest possible length.
- *Usual assumptions:*
  - we cover a rectangular area;
  - each on-board sensor covers all the points within a given radius r.
- *What we do:* describe the trajectories which are (asymptotically) optimal under these assumptions.

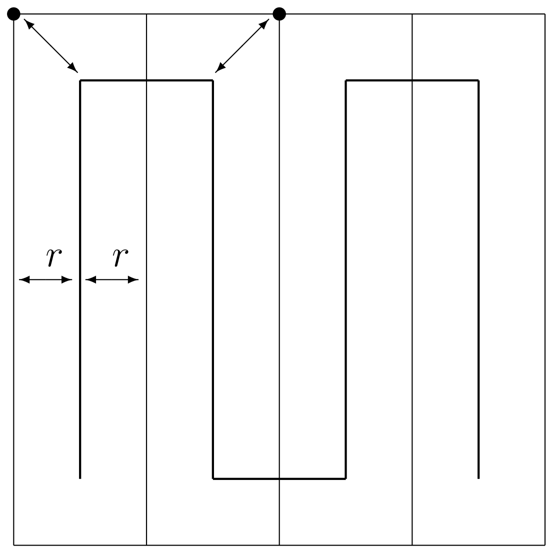
18. An (Almost) Optimal Trajectory



- In the region of area $A_0 = L_1 \cdot L_2$, we have $\dfrac{L_1}{2r}$ pieces of length $\approx L_2$ each.
- The total length is $L \approx \dfrac{L_1}{2r} \cdot L_2 = \dfrac{L_1 \cdot L_2}{2r} = \dfrac{A_0}{2r}$, i.e., this trajectory is (almost) optimal.

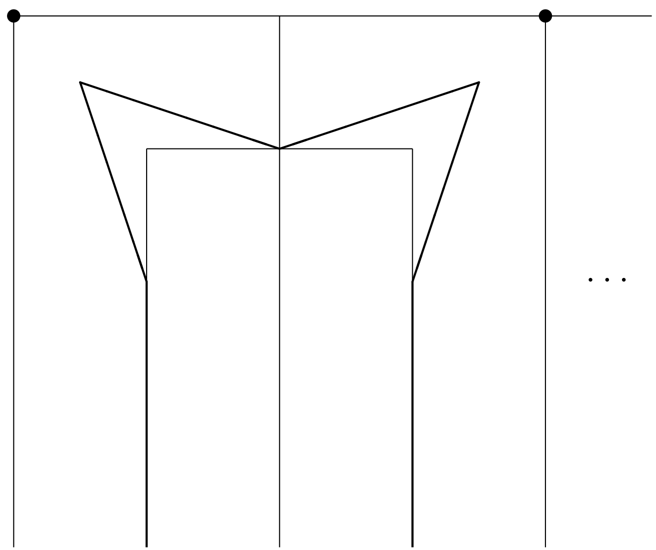
Home Page
Title Page
◀◀ ▶▶
◀ ▶
Page 20 of 24
Go Back
Full Screen
Close
Quit

19. Minor Problem



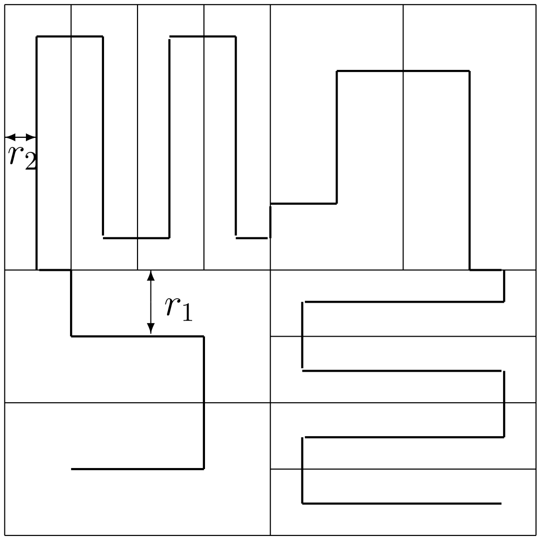
- *Problem:* corner points (marked bold) are not covered.
- *Explanation:* the distance from the trajectory to each corner point is $\sqrt{r^2 + r^2} = \sqrt{2} \cdot r > r$.

20. Solution: How to Cover Corner Points



- *Comment:* this way, corner points are covered.

21. What If We Want Different Coverage In Different Sub-Regions: Asymptotically Optimal Solution



- *Idea:* use (asymptotically optimal) arrangement in each sub-region; this sub-division can be iterated.

22. What If We Want Different Coverage In Different Sub-Regions: General Case