

**נושא הקורס בתשפ"א:** הסתברות בממד גבוה ויישומה במדע הנתונים

**שם עברי מקוצר:** הסתברות בממד גבוה

**דרישות קדם:** אותות אקראיים (044202). קורס בתורת האינפורמציה (046733) מומלץ אבל לא מחייב. נדרש ידע בסיסי באנליזה פונקציונלית.

**סילבוס בעברית:** תחומים מדעיים ויישומיים רבים הרחיבו לאחרונה את מעטפת היכולות שלהם הודות לאיסוף נתונים נרחב. המתדולוגיה של הסתברות בממד גבוה וניתוח לא אסימפטוטי מאפשרת הבנה יסודית של האפשרויות למיצוי מידע זה, וכוח מניע לשיפור ביצועים. הקורס יציג את העקרונות והכלים של תיאוריה זאת, וידגים את יישומיהם במגוון בעיות הסקה ולמידה סטטיסטית.

### **נושאי לימוד:**

ילמדו הנושאים הבאים:

1. מבוא להסתברות בממד גבוה וסטטיסטיקה לא-אסימפטוטית.
2. משתנים אקראיים תת-גאוסיים ואי-שיוון Hoeffding. משתנים תת-מעריכיים ואי-שיוון Bernstein. מרחבי Orlicz של משתנים אקראיים. יישומים בבעיות שערך וסיווג.
3. אי-שיוויונות מקסימליים, רשתות וחסמים על מספרי כיסוי ואריזה. יישומים בבעיות גרסיה לינארית כללית ובבעיות עם תנאי דלילות.
4. ביצועי מינימקס בהסקה סטטיסטית: תכונות בסיסיות ותכונות טנזוריזציה של מדדי שונות-אינפורמציונית. שיטת שתי ההשערות של Le-Cam והרחבה למספר השערות בשיטת Fano.

בהתאם לזמן ולעניין, יועברו הנושאים הבאים:

5. אי-שיוויוני ריכוז של סכום מטריצות בלתי-תלויות ואי-שיוון Bernstein מטריצי. ריכוז של אופרטורים של מטריצות אקראיות ומשפט Davis-Kahan. יישומים בבעיות של שערך מטריצה, ביטול-רעש, זיהוי קהילות ברשתות, וניתוח רכיבים עיקריים (PCA).
6. תהליכים אקראיים גאוסיים, ואי-שיווני השוואה של Slepian, Sudakov-Fernique. מינורציה של Sudakov.
7. סימטריזציה וביטול-צימוד. שיטת השרשור של Dudley. חוק מספרים גדולים במידה-שווה. קשרים בין מספרי כיסוי לממד Vapnik-Chervonenkis -- הלמה של Sauer-Shelah והלמה של Dudley. חסמי הכללה בבעיות של למידה סטטיסטית.

### **מקורות:**

1. High Dimensional Probability: An Introduction with Applications in Data Science, by R. Vershynin, 2019.
2. High-Dimensional Statistics: A Non-Asymptotic Viewpoint, by M. J. Wainwright 2019.

3. High Dimensional Statistics, Lecture notes by P. Rigollet and J-C. Hutter, 2017.
4. Information-theoretic Methods for High-dimensional Statistics, Lecture notes by Y. Wu (Yale), 2019.
5. Probability in High Dimension, Lecture notes by R. van Handel (Princeton), 2016.

### **תוצאות למידה:**

הסטודנטים יוכלו לנסח, להכליל ולעדן מודלים מתמטיים לתרחישי ניתוח נתונים, סטטיטיקה בממד גבוה ולמידה סטטיסטית, להפעיל כלים של הסתברות בממד גבוה לניתוח שלהם ולתרום לחזית המחקר המדעי בנושא זה. הסטודנטים יכירו את המאפיינים הייחודיים של גדלים אקראיים בממד גבוה.

**הרכב הציון:** 30% תרגילי בית, 70% עבודת גמר.

### **שם הקורס באנגלית:** Probability in high dimension

**English syllabus:** Numerous scientific fields have recently expanded their capabilities thanks to extensive data collection. The high-dimensional probability methodology enables an understanding of the fundamental limits of extracting information from such data. The course will present its elements, and demonstrate its applicability in statistical inference and learning problems.

### **Topics:**

1. Introduction to probability in high dimension and non-asymptotic statistics.
2. SubGaussian random variables and Hoeffding's inequality. SubExponential random variables and Bernstein's inequality. Bernstein's conditions. Orlicz spaces. Applications in estimation and classification problems.
3. Maximal inequalities. Nets, covering and packing numbers. Applications in unconstrained linear regression and under sparsity assumptions.
4. Concentration of matrix norms. Concentration of sums of independent matrices, and matrix Bernstein's inequality. Davis-Kahan inequality. Applications in matrix estimation, matrix denoising, community detection, and principal component analysis (PCA).
5. Minimax lower bounds. Basic and tensorization properties of information divergences. Le-Cam's two point method and multiple-hypotheses Fano's method.
6. Gaussian random processes, and Slepian's comparison inequality. Sudakov-Fernique comparison and Sudakov minorization.
7. Decoupling and symmetrization. Dudley's chaining integral and uniform laws of large numbers. Connections between covering numbers and Vapnik-Chervonenkis dimension – the lemmas of Sauer-Shelah and Dudley. Generalization bounds in statistical learning.

**Learning Outcomes:** The students will be able to formulate, generalize and refine mathematical models for data-science scenarios, high-dimensional statistics, and statistical learning. They will be able to utilize tools from high-dimensional probability to analyze them, and contribute to the forefront of research in these topics. The students will be acquainted with the unique characteristics of random structures in high dimension.