This course is about an approach to designing algorithms for machine learning problems by searching for sets of values that look like the moments of a set of random variables that satisfy a given family of bounds and constraints. For example, I might assert that $X$ is Boolean and $Y$ takes values in the range $[0,10]$, and $\mathbb{E}[XY]=5$. Then if $X=1$ and $Y=5$, this is a trivial distin’ that is consistent with these requirements, with moments $\mathbb{E}[X]=1, \mathbb{E}[Y]=5, \mathbb{E}[X^2]=1, \mathbb{E}[Y^2]=25$. But, we could also have $X$ as a $\frac{3}{4}$-biased coin, and $Y$ independently takes value 10 w.p. $\frac{2}{3}$, and 0 w.p. $\frac{1}{3}$. Then indeed, $\mathbb{E}[XY]=\frac{1}{4} \cdot 0 + \frac{3}{4} \cdot \frac{1}{3} \cdot 0 + \frac{3}{4} \cdot 2 \cdot 3 \cdot 10 = 5$, and here $\mathbb{E}[X]=\frac{3}{4}, \mathbb{E}[Y]=\frac{20}{3}, \mathbb{E}[X^2]=\frac{3}{4}, \mathbb{E}[Y^2]=\frac{200}{3}$, … Both of these, and many other sets of values would be solutions to this set of bounds and constraints. This is an example of a moment problem.

For us, moment problems will be given by systems of polynomial inequalities. For a polynomial $p(x)$, we have bounds of the form $\mathbb{E}[p(X)] \geq 0$ or constraints $p(X) \geq 0$. We then want to find values $\mathbb{E}[X^k]$ (e.g. $\mathbb{E}[X], \mathbb{E}[X^2], \ldots$).
\( E[x^2], E[x_1 x_2], \ldots, E[x^3], \ldots \) for a probability distribution consistent with these bounds and constraints, or report that no such distribution can exist. E.g., suppose \( X_{\text{level}} \geq 0, \Pr[\text{high} X] (= E[\text{high} X]) \geq \frac{1}{10} \) where high \( X \) is Boolean \( \Leftrightarrow \text{high} X^2 - \text{high} X = 0, \text{high} X = [X_{\text{level}} \geq 10] \) \( \Rightarrow (X_{\text{level}} - 10) \text{high} X \geq 0 \) and \((10 - X_{\text{level}}) (1 - \text{high} X) \geq 0\), but we also want \( E[X_{\text{level}}] \leq \frac{1}{2} \). Notice, \( E[X_{\text{level}}] \geq E[X_{\text{level}} | \text{high} X] \Pr[\text{high} X] + E[X_{\text{level}} | \text{not high} X] (1 - \Pr[\text{high} X]) \geq 10 \cdot \frac{1}{10} + 0 = 1 > \frac{1}{2} \). So there's no solution to this system. What we'll actually be finding are sets of values that satisfy all of the bounds like the above, that we could derive, when possible.

Actually optimizing over the sets of possible moments is intractable in general, so we solve an easier problem described by a "semidefinite program."

**Semidefinite Programming**

A semidefinite program is an optimization problem, like a linear program (we have linear constraints on real valued vectors of variables), in which we may additionally assert that a matrix of linear forms is positive semidefinite: recall, \( M \) is positive semidefinite if for all real vectors \( \bar{x} \), \( \bar{x}^T M \bar{x} \geq 0 \). Equivalently for symmetric \( M \), if \( M \) has a Cholesky factorization \( M = U^T U \) for some \( U \). (Indeed, \( \bar{x}^T U^T U \bar{x} = (U \bar{x})^2 \geq 0 \) for all \( \bar{x} \); and conversely, recalling
that we can obtain eigenvectors of $M$ by maximizing the Rayleigh quotient $\frac{\mathbf{v}^T M \mathbf{v}}{\|\mathbf{v}\|^2}$ in the subspace orthogonal to $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{i-1}$. We see that the PSD constraint ensures that the eigenvalues are all nonnegative—so, taking $U = [\mathbf{v}_i \mathbf{v}_j] \mathbf{V}$ gives $U^T U = \mathbf{V}^T [\mathbf{v}_i \mathbf{v}_j] [\mathbf{v}_i \mathbf{v}_j] \mathbf{V} = M$. Indeed, non-negativity of the eigenvalues is another equivalent characterization.)

There exist polynomial-time algorithms for solving such systems of constraints up to a polynomial number of bits of precision, provided that there exists a solution that can be represented in a polynomial number of digits (i.e., has size $\leq 2^{\mathsf{poly}(n)}$). Some semidefinite programs have only large solutions, so we need to take some care.

We will be writing semidefinite programs to approximately capture our moment problems as follows. First, consider the following “moment matrix” for a given degree $d$: letting $\mathbf{x}^{\leq d/2}$ be the vector of monomials of total degree up to $d/2$, e.g., $(x, y)^{\leq 2} = (1, x, y, x^2, xy, y^2)$, the moment matrix is $\mathbb{E}[\mathbf{x}^{\leq d/2} (\mathbf{x}^{\leq d/2})^T]$—e.g.,

$$
\begin{bmatrix}
1 & x & x^2 & xy & y^2 \\
x & x^2 & xy & x^2y & y^2 \\
x y & xy & x^2y & xy^2 & y^2 \\
x^2 & x^2y & x^3 & x^2y^2 & x^2y^2 \\
x y^2 & xy^2 & x^2y^2 & xy^3 & x^2y^2 \\
y^2 & xy^2 & x^2y^2 & xy^3 & y^3 \\
x^2 & x^2y^2 & x^3 & x^2y^2 & y^3 \\
x y^2 & xy^2 & x^2y^2 & xy^3 & y^3 \\
y^2 & xy^2 & x^2y^2 & xy^3 & y^3 \\
x^2 & x^2y^2 & x^3 & x^2y^2 & y^3 \\
x y^2 & xy^2 & x^2y^2 & xy^3 & y^3 \\
y^2 & xy^2 & x^2y^2 & xy^3 & y^3
\end{bmatrix}
$$

Notice—this matrix is indeed, for any vector $\mathbf{p}$, $\mathbf{p}^T \mathbb{E}[\mathbf{x}^{\leq d/2} (\mathbf{x}^{\leq d/2})^T] \mathbf{p} = \mathbb{E}[(\mathbf{p}^T \mathbf{x}^{\leq d/2})^2] \geq 0$. Likewise, supposing we know $q(\mathbf{x}) \leq \mathbf{q}^T \mathbf{x} \geq 0$ holds for the entire distribution, then the matrix $\mathbb{E}[(\mathbf{q}^T \mathbf{x}) \mathbf{x}^{\leq d/2} (\mathbf{x}^{\leq d/2})^T]$ is also positive semidefinite.
where both \( q^* x \pm \left( p^* x^{d/2} \right)^2 \) are nonnegative. We will put \( d' = d - \deg(p) \) so that the resulting expression only has monomials of total degree at most \( d \). This is the degree-\( d \) localizing matrix for the constraint \( q(x) \geq 0 \). If we had a constraint \( q(x) = 0 \), then clearly \( \mathbb{E}[\tilde{q}^* x (x^{d/2} (x^{d/2})^T) \tilde{q}] = 0 \), and this is the localizing matrix for \( q(x) = 0 \). Now, for a given system of bounds \( \mathbb{E}[b_i(x)] \geq 0 \) and constraints \( q_j(x) \geq 0, h_k(x) = 0 \) the degree-\( d \) sum of squares relaxation is the following semidefinite program: the degree-\( d \) moment matrix is \( \text{psd} \), \( \mathbb{E}[b_i x^{d}] \geq 0 \) for each bound (a 1x1 \( \text{psd} \) constraint), the degree-\( d \) localizing matrix for each \( g_j(x) \geq 0 \) is \( \text{psd} \), and each entry of the degree-\( d \) localizing matrix for each \( h_k(x) = 0 \) is equal to 0, i.e., \( \mathbb{E}[h_k x^2] = 0 \) where \( x^2 \) ranges over monomials of total degree \( \leq d - \deg(h) \).

We have argued that the moments of any set of random variables satisfying the bounds and constraints are feasible solutions to this semidefinite program, but they may not be the only solutions. What we can say about them is that they satisfy all of the bounds that can be proved using “sum-of-squares” proofs. We'll see that these are surprisingly rich, so solutions to the programs are adequate for many things.

**Sum of squares proofs**

A sum-of-squares proof is a way of establishing that a system of bounds \( \mathbb{E}[b_i(x)] \geq 0 \) and constraints \( q(x) \geq 0, h_k(x) = 0 \)
has no solution. It is an expression of the form
\[ \sigma_0(x) + \sum_i \lambda_i b_i(x) + \sum_j \sigma_j(x) g_j(x) + \sum_k q_k(x) h_k(x) \quad \text{formally} \quad -1 \]
where \( \sigma_i(x) \)'s are sum-of-squares polynomials \( (= \sum_k p_k(x)^2) \lambda_i \in \mathbb{R}^{\mathbb{R}_+} \), and \( q_k(x) \)'s are polynomials. In particular, if every polynomial expression (before cancellation) has total degree at most \( d \), we say that it is a degree- \( d \) sum-of-squares refutation. Indeed, it establishes infeasibility since \( E[\sigma_0(x)] \geq 0, E[\sum_i \lambda_i b_i(x)] \geq 0, E[\sum_j \sigma_j(x) g_j(x)] \geq 0 \) and \( E[\sum_k q_k(x) h_k(x)] = 0 \) by assumption, so the LHS \( \geq 0 > -1 = E[-1] \) which is the RHS.

But conversely, if the system is "explicitly bounded"—i.e., if there exists \( \sigma_0(x) \) SOS and \( M > 0 \) such that \( \sum x_i^2 + \sigma_0(x) = M \), (this will be true for the systems we consider...) then whenever the degree- \( d \) sum-of-squares relaxation is infeasible there is a degree- \( d \) sum-of-squares refutation.

Let's see this by the contrapositive—we'll show that if no sum-of-squares refutation exists, then there is a solution to the semidefinite program.

First recall that a cone \( \mathcal{C} \) is any set closed under nonnegative linear combinations: \( x \in \mathcal{C}, y \in \mathcal{C} \Rightarrow \alpha x + \beta y \in \mathcal{C} \quad \forall \alpha, \beta \geq 0 \).

Thus, cones are convex. Observe that sum-of-squares expressions form a cone. If there is no sum-of-squares refutation, this means that \(-1\) (the polynomial) does not lie in this cone. Claim: \(-1\) does not lie on the boundary, either. Therefore, by the separating hyperplane theorem, there is a
linear form $L$ s.t. $L(-1) < 0$ but $L(p(x)) \geq 0$ for all poly
omials $p(x)$ in the cone, in particular, if we put $\mathbb{E}[x^2] = L(x^2)$,
since $b(x)$, $p(x) = -\frac{p^* x \cdot \hat{d}x}{\hat{d}x \cdot \hat{d}x} \hat{d}$, $g_j(x) p(x) = -\frac{p^* g_j(x) x \cdot \hat{d}x}{\hat{d}x \cdot \hat{d}x} \hat{d}$,
$h_k(x) x^2$ and $-h_k(x) x^2$ are all in the cone, the resulting $\mathbb{E}[x^2]
$ satisfies all of the constraints of our SDP. (We may just need to rescale by $L(1)$ to get $\mathbb{E}[1] = 1$.)

Proof of claim: By contrapositive: suppose $-1$ lies on the boun-
dary, i.e., there is a $p(x)$ in the cone s.t. $-1 + \varepsilon p(x)$ also lies in
the cone for any $\varepsilon > 0$. Now, by explicit boundedness, we
find that for some $\sigma'(x)$ SOS and $M' > 0$ that $M' - p(x) - \sigma'(x)$
is in the cone as well. (That is, $p(x) \leq M'$ e.g., to show $x_i \leq \sqrt{M}
from \Sigma x_j^2 \leq M$ i.e., get $x_i + \sqrt{M}$ from $M - \Sigma x_j^2$ use $\frac{1}{2 \sqrt{M}} (M - \Sigma x_j^2) + \frac{1}{2 \sqrt{M}} (x_i + \sqrt{M})^2 + \frac{1}{2 \sqrt{M}} (x_i + \sqrt{M})^2$
We can build up $M' - p(x)$ term-by-term similarly.)

But now, for $\varepsilon < \frac{1}{M'}$, $-1 + \varepsilon p(x) + \varepsilon M' - \varepsilon p(x) - \sigma'(x) + \varepsilon \sigma'(x) = \varepsilon M' - 1
is in the cone as well. Rescaling by $\frac{1}{\varepsilon M'}$ (note: $\varepsilon M' < 1$) gives $-1$ is
also in the cone.

I stress that the SDP was actually infeasible earlier: note that $\mathbb{E}[(w(x)) = \mathbb{E}[\sigma_0 \Sigma x \cdot \hat{d}x \cdot \hat{d}x] \geq 0$ for any vector $\sigma_0
since $\mathbb{E}[x^{\hat{d}x} x^{\hat{d}x}] \geq 0$ in any feasible solution, and similar-
ly for the other constraints. Thus indeed, a solution to
the SDP and the existence of a sum-of-squares refutation
are mutually exclusive, and one must exist (for explicitly
bounded systems). But now, e.g., if we can derive $p(x) \leq M
with a sum-of-squares proof ($M - p(x)$), then adding a bound
$E[p(x) - M - \delta]$ gives an infeasible system: \( \frac{1}{\delta}(M - p(x)) + \frac{1}{\delta}(p(x) - M - \delta) = -1 \)

Thus, any solution to the SDP must obey \( E[p(x)] \leq M \).

In this way, solutions to the SDP obey all bounds with Sos proofs.