The Bayesian interpretation of pseudo-distributions.

Consider a an optimization problem of the form $\min_{x\in\{0,1\}^n} f(x)$. When we run the degree d sos algorithm on this problem, we obtain the minimum value $\alpha \in \mathbb{R}$ for which there is a degree d pseudo-distribution μ such that $\tilde{\mathbb{E}}_\mu f = \alpha$. Our notation strongly encourages us to pretend that μ is an *actual* distribution over $x \in \{0,1\}^n$ such that $f(x) = \alpha$.[1] Indeed, many properties of pseudo-distributions, such as satisfying the Cauchy-Schwarz and Holder inequalities, are most naturally explained by the phenomena of "inheriting" traits of actual distributions. However, often we have strong reasons to believe that no such distribution exists, or that if it does, its moments look nothing like μ's. How are we to interpret μ in such cases?

[1] Satisfying $\tilde{\mathbb{E}} f = \alpha$ and being supported on $\{x : f(x) = \alpha\}$ is not the same thing in general, but these notions do coincide when α is the global minimum of f.

Let us elaborate on this a bit more. In many natural settings, the *global minimum* of f would be achieved at a single point. For example, if we think of f as counting the number of violations of some constraints or equations, then often these equations are *over constrained* in the sense that the number of constraints is large enough to completely determine the optimum point. In optimization problems arising from machine learning, $x^* \in \{0,1\}^n$ could denote the unknown parameters of the model, and f would be a function derived from the observed data.[2] As we accumulate enough data, eventually it would completely determine the unknown parameters (with arbitrarily high accuracy) in a statistical sense. So, if μ was truly supported on points achieving the global minimum then μ would simply be the distribution putting all the weight on the single point x^* which in particular means that $\tilde{\mathbb{E}}_\mu x_i = x^*_i$ for every i. But if deriving the parameters from the observations is *computationally hard* then we shouldn't be able to "read off" x^* from the expectations and so it will *not* hold that $\tilde{\mathbb{E}}_\mu x_i$ is equal to either 0 or 1 for all i.

[2] E.g., x^* can denote the parameters of an unknown neural network and f is the function that minimizes the total deviation of the predictions of the candidate network x from the correct labels on a large number of samples.

As another example, consider the case where $f(x)$ counts the number of violations of the constraints of some particular 3SAT formula φ by the assignment $x \in \{0,1\}^n$. The forumla φ is *satisfiable* if and only $\min_x f(x) = 0$ but since this problem is NP hard there should exist some f's of this form such that $\min_x f(x) > 0$ but for every constant d there is a d pseudo-distribution μ such that $\tilde{\mathbb{E}}_\mu f = 0$. Thus this pseudo-distribution pretends to distribution over elements (i.e., satisfying assignments) *that don't exist*. This is why we sometimes refer to such pseudo-distributions as akin to being supported over "unicorns".

Bayesian probability theory

> *"The probability of winning a battle" . . . has no place in our theory . . . because we cannot think of a collective to which it belongs. The theory of probability cannot be applied to this problem any more than the physical concept of work can be applied to . . . the "work" done by an actor reciting his part.* Richard Von Mises, 1928

> *The theory of inverse probability is founded upon an error, and must be wholly rejected*, Fisher, 1925

> *If anyone wishes to study the properties of frequencies in random experiements he is, of course, perfectly free to do so; and we wish him every success. But . . . why does he insist on appropriating the word "probability" which had already a long-established and very different technical meaning?* E.T. Jaynes, 1978

> *I am unable to see why requires us to interpret every probability as a frequency in some random experiment; particulary when . . . probabilities appearing in most problems are . . . frequencies only in an . . . imaginary universe invented just for the purpose of allowing a frequency interpretation.*, E.T. Jaynes, 1976

> *"The terms and describe the various degrees of rational beliefs about a proposition . . . All propositions are true or false but the knowledge we have of them depends on our circumstances."* John Maynard Keyenes, 1921

What kind of a probability theory would allow you to assign a value different than 0 or 1 to an event that clearly either happened or not? It turns out that this is related to a longstanding question in the philosophy of probability and in particular to the debate between the *Bayesian* and *Frequentist* interpretation of probability.[3] Probability theory initially arose in the 17th century from Pascal's study of *gambling*. The concern there was how to make the best possible bets. Making bets (like investing in stocks) is less about the inherent uncertainty of the outcome and more about our *information* about this outcome: different bettors have different knowledge about the horse, team, company, or whatever we want to bet on. Indeed probability theory, as was studied before the 19th century, was mostly based on the viewpoint of the *observer*, which generally can be subjective but in some cases, such as tossing a die, essentially all observers have the same information. That is, for all observers all six sides are *symmetric* which suggests assigning a probability of 1/6 for each particular outcome. This idea, which is known as the Principle of Indifference is at the cornerstone of classical probability theory as originated by Bernoulli and Laplace and allowed them to argue even about the probabilities of events that do not come from a well defined sample space such "the probability that the sun will rise tomorrow".

However in the late 19th and 20th century, researchers such as Fisher, Neyman and Pearson found the symmetry/subjective based

[3] Obviously we cannot do justice to this deep issue in these lecture notes. Two classic books on this topic from the frequentist and Bayesian viewpoint respectively are Feller [1968] and Jaynes [2003] . This talk by Michael Jordan is also a good starting point for this discussion.

approach to probability lacking in rigor, and have moved to the sample space or *frequentist* approach which is how we typically learn probability in mathematics today.[4]

In this approach, a probability distribution is defined by a function μ that assigns to any element x in some *sample space* S some probability $\mu(x) \in [0,1]$ (where these numbers sum, or integrate, to 1), and the probability of some event A is obtained by summing up $\mu(x)$ for all $x \in A$.

The frequentist viewpoint is particularly well suited for *hypothesis testing*. This is the setting where we have some *hypothesis* H_0 (known as the "null hypothesis") and can set up an experiment that corresponds to a sample space in which if H_0 is true then the probability that some event A occurs is at most $1/2$. If we then repeat the experiment k times, and in all the cases the event A occurs, then we can decide that H_0 has been "refuted" and we would be wrong at most a 2^{-k} fraction of the times. The value 2^{-k} is known as the "p-Value" and in many scientific settings it is set to be 0.05. For example, if we are given a coin and our hypothesis is that it is a fair coin, then if we toss it 5 times and check if we got all heads then, if so, we can conclude that it is not a fair coin with a p-value of $1/2^5 = 1/32 \sim 0.03$.[5] This means that if we use this as our standard test for fairness of coins then the *frequency* in which we would make an error calling a coin unfair when it is fair would be at most $1/32$.

Indeed, for *frequentist* statisticians, probability does not make sense if we can't set up a potentially repeatable experiment with well defined sample space. Note that the frequentist outlook accepts the fact that sometimes (indeed up to a p fraction of the times, even if other sources errors are accounted for) we will come to the wrong conclusion. The frequentist outlook is well suited for experiment design in science, in the sense that it is *not* about using all available information to deduce whether or not the hypothesis is true, but rather about designing an experiment that can be rigorously analyzed to give evidence about its truth. If we had to bet on the hypothesis' truth, we might want to try to use more of the information to get the best possible estimate for its likelihood, but the frequentist approach is not designed to achieve a betting strategy to make the best possible guess using the given data. Rather it is about designing an experiment that would gather more data until we can make a high confidence prediction.[6]

Not all situations in which we want to apply probability, whether it is betting, investing, or making decisions, fall cleanly into the frequentist framework. Sometimes we do need to make some predic-

[4] In particular the application of probability theory to events that are not repeatable experiments can *sometimes* lead to somewhat unsettling results such as the so called "doomsday argument" by which one argues that humanity is likely not to survive in similar form for an extended period of time. This is because about 100 billion people have ever born so far and if you assume a uniform prior on the person you happened to be among all the people that ever existed or will exist, then this latter total should not be much more than 200 billion.

[5] This example also shows the importance of deciding on the test to perform in advance: clearly *any* particular outcome of the five coin tosses has probability 2^{-5} and so the mere fact that the outcome was "unlikely" is no evidence that the coin wasn't fair.

[6] One analogy is to consider the problem of trying to decide on the merits of a legal case. The Bayesian approach would be to take all the possible information and beliefs we have access to into account and then make the best possible prediction. The Frequentist approach is to define rigorous "rules of evidence" that may sometimes not allow us to derive the truth, but can be shown to be correct most of the times. One can see that the Bayesian approach is more appropriate if we are interested in making a subjective judgement in a "one off" case that is as accurate as can be given our prior knowledge and the data we have, while the frequentist approach is better suited for deciding on procedures that will allow us to reach some mutually agreed upon objective conclusions in *many* cases.

tions outside a controlled experimental environment, and we want to do the best job we can with the data we have. These kind of "messy" situations are occurring more often in modern applications and have led to a sort of partial resurgence of the "subjective"/"betting" view of probability theory, which is now commonly known as *Bayesian* probabilities. The reason for the name *Bayesian* is that such calculations almost always rely on Bayes' theorem which states that if you made some observation A, then the probability you should assign to an unknown B should conditioned on this observation should be

$$\mathbb{P}[A \cap B] / \mathbb{P}[A] \tag{1}$$

In the context of subjective or betting probabilities, one initially has a *prior* distribution μ over the set of possible outcomes, and then when we learn that some event A occurred and we need to make a bet on B then we use Eq. (1) to derive the odds.

Note that in the context of betting, it may make perfect sense to give odds that do not correspond to zero or one probabilities even to events that have been fully determined. For example, if I had to bet on whether my great great grandfather had blue eyes, I would probably try to estimate this probability based on some prior estimate for the prevalence of blue eyes, adjusting it based on observations of the eye color of his descendants. However, clearly he either had blue eyes or didn't, and so this probability is really due to my ignorance rather than to any inherent unpredictability of this event. More commonly, whenever we assign probabilities to events such that "X will win the upcoming election" or "company Y will perform well in the stock market", there isn't any well defined sample space out of which this event is drawn out of repeated experiments. Rather the way to interpret these probabilities is that these represent the best odds we can give to those events given the information at our disposal.

A Bayesian and a frequentist go to the race tracks

The key philosophical difference between Bayesians and Frequentists is whether the probability of an event A correspond to the *beliefs* of an observer on the likelihood of A or to the frequency that A would occur in repeated identical but independent experiments. What is more interesting to us is how this philosophical is manifested mathematically.

Suppose that Betty the Bayesian and Frida the frequentist go to the horse race tracks where they can bet on various outcomes of a horse

race (who will come first, what time it will take, the gap between first and second place etc. etc.). We can model the outcome of such a race as a random variable X that is sampled from some distribution $p(X|\theta)$ where θ is some set of (unknown) parameters on the inherent abilities of the horses, jockeys, etc.. Betty and Frida observe various partial information (e.g., results of past races, which we can think of as some random variable Y correlated with X) and then need to make bets on it. We can think of a "bet" as some function h that maps X to $\mathbb{R}$, where $h(x)$ is the payout of the bet when the outcome of X is x. Let's think of the bettors goal is to come up with *prices* $\Phi_{Betty}(h)$ and $\Phi_{Frida}(h)$ for every potential bet h such that Betty (resp. Frida) would be willing to either buy or sell the bet h at the price $\Phi_{Betty}(h)$ (resp. $\Phi_{Frida}(h)$). One can see that the best price (that is guaranteed not to lose in expectation regardless if you buy or sell) for the bet h would be $\mathbb{E}_{x\sim p(X|\theta)} h(x)$ but the problem is that we don't know θ.

Betty would handle this as follows: she will try to encode all her prior knowledge and assumptions on the ability of the horses into a *prior* distribution $\mathcal{D}$ on the possible values of θ.[7] Then she updates this prior based on the observations y to obtain the *posterior distribution* $\mathcal{D}' = \mathcal{D}|Y = y$. Betty's price $\Phi_{Betty}(h)$ is the expectation of $h(x)$ when we sample θ conditioned on y and then x conditioned on θ. Betty's price is only as good as the prior she uses, and we have no guarantee on what her performance would be for arbitrary θ. That is, there is no no guarantee of **calibration** betweeh $\Phi_{Betty}(h)$ and the empirical average of $h(x)$ if we were to sample x many times from $p(X|\theta)$.[8] But it least Betty's prices are **coherent** in the sense that they correspond to *some* distribution over the x's. In particular if for every x, $h(x) \geq h'(x)$ then we are guarantees that Betty's price will for h will always be at least as large as her price for h', and so there is no "dutch book" or "arbitrage" strategy against Betty that is guaranteed to make money off her regardless of what the value of θ is.

For Frida, the parameters θ and the random variable X play very different roles. The choice of θ already happened and so from a frequentist viewpoint, there is no sense in assigning it a probability distribution. Rather Frida will try to come up with some estimate $\Phi_{Frida}(h)$ that is guaranteed to have some **calibration** in the sense of a bound on the difference between $\Phi_{Frida}(h)$ and $\mathbb{E}_{x\sim p(X|\theta)} h(x)$ that holds for *every* θ from a set Θ of potentially allowable θ's and such that if we can think of the data Y as coming from independent measurements $Y_1, \ldots, Y_n$ then as n goes to infinity this difference tends to zero. Unlike in the Bayes case, there is no universal prescription of how to come up with the frequentist estimate, and so Frida's

[7] The reliance on a subjective prior might make us uncomfortable, but note that in most cases the choice of the model for X (i.e., the distribution $p(X|\theta)$) is already subjective as well. In practice people often do the equivalent of "looking for a coin under the streetlight" and change the model so it becomes more computationally tractable.

[8] Betty's prices do satisfy *average case calibration* in the sense that if θ is sampled from Betty's prior $\mathcal{D}$ and then x, y are sampled conditioned on θ then the expectation of $\Phi_{Betty}(h)$ (which is a function of y) would be the same as the expectation of $h(x)$.

focus is on getting some estimation procedure that she can *analyze* to rigorously guarantee the calibration property. Since Frida's estimation procedure might not correspond to taking the expectation of h under any distribution, it might be **incoherent** and in praticular there might be two functions h, h' such that $h(x) \geq h'(x)$ for every x but $\Phi_{Frida}(h) < \Phi_{Frida}(h')$. So there could be a "dutch book"/"arbitrarge strategy" against Frida where we would buy h bets from her and sell her h' bets and be guaranteed to make money regardless of the value of θ.

The discussion so far ignored the question of *efficiency*. For Frida requiring the map $h \mapsto \Phi_{Frida}(h)$ to be efficient restricts the set of estimation procedures that she can use. For Betty, this is more complicated since this restriction can (and often does) rule out the unique estimation procedure. In such a case, typically Frida would compute her estimate using an efficiently sampleable distribution $\mathcal{D}'$ that is not the true posterior $\mathcal{D}|Y = y$ but is hopefully somewhat related to it. She could obtain $\mathcal{D}'$ by simplifying the model or her prior (i.e., in effect "forgetting" some information, such as higher order correlations, that is hard to incorporate in a computationally efficient manner), or run a random walk type algorithm that would sample from the true posterior $\mathcal{D}$ in the limit, but stop it before it does so. This may make her estimate even less calibrated than it was before, but, since it comes from a probability distribution, it would still be *coherent*. While for inefficient algorithms, restricting attention to estimation procedures that are based on actual distributions is without loss of generality, as we'll see, in the efficient case this can come at a significant cost.[9] In particular, as we will see, it is possible that no sampleable distributions will match the observations y, and hence there, by observing y, we could have a "dutch book"/"arbitrage" strategy that makes money off Betty without any risk.

[9] This is similar, and related to, the fact that in mechanism design in economics by the revelation principle we can always have a *truthful* mechanism if we don't care about efficiency, but for *efficient* mechanisms this is not necessarily the case **?** .

The *sum of squares* algorithm can be thought of as some "hybrid" between the frequentist and Bayesian approach. Like in the Bayesian case, the sos algorithm is a single algorithm, and we can write an sos program such that the pseudo-distributions solving it would correspond to the BAyesian procedure if the degree d goes to infinity.[10] However, when we want to have an efficient algorithm, then sos turns into a frequentist procedure, since rather than coming up with an actual probability distribution $\mathcal{D}'$ that "approximately" matches the observations, we come up with a *pseudo distribution* that *exactly* matches them. Since it is a pseudo-distribution, it does not (and generally will not) satisfy the coherence property, but it would satisfy **degree** d **pseudo coherence** in the sense that if there is a degree d proof that $h \geq h'$ then our price for h would always be at least

[10] The key notion here is that since the set of pseudo-distributions over pairs (θ, x) that satisfy the observations y is *convex*, we can minimize over it the convex function corresponding to a distance from our prior.

as large as the price for h'. Moreover, it will also satisfy a **degree** d **pseudo calibration** condition in the sense that if we have a degree d SOS proof that conditioned on y, for every θ, $h(x) \in [\alpha, \beta]$ then the estimate will also fall in this interval. We now elaborate on this more, using another example: that of the *clique* problem.

Contrasting sos and Bayesian analysis: an example

Consider the following: We are given a graph G on n vertices that represents some data we have gathered. For example, the vertices of G might be certain proteins and edges correspond to *interactions* between them that have been observed in patients with disease X. Suppose that we posit that the underlying mechanism of the disease is captured by a set of k proteins that all interact with one another. That is, a *clique* in the graph G, which we represent as a vector $x \in \{0,1\}^n$ such that $\sum x_i = k$ and $x_i x_j = 0$ if i is not connected to j.

Now suppose that we have N different candidate drugs, each of which will either cure the disease or not depending on the underlying mechanism. So we can think of these drugs as N functions $f_1, \ldots, f_N \colon \{0,1\}^n \to \{0,1\}$ where $f_\ell(x) = 1$ if and only if drug ℓ will cure the disease if its mechanism is x. We now want to understand which drug is most worthwhile performing an experiment on.

In the classical Bayesian approach we assume that there is some underlying *prior* distribution p on pairs (G, x) where x is a k-clique in G.[11] For example, we can take the "maximum entropy" prior that x is a uniformly random vector in $\{0,1\}^n$ of weight k, and G is a random Erdos-Renyi graph where x is "planted" as a clique. Once we observe G, we then define the *posterior* distribution $p(x|G) = p(x,G)/p(G)$ using Bayes' law. Now we can define the expected utility of drug ℓ as $\mathbb{E}_{x \sim p(x|G)} f_\ell(x)$.

[11] In the notation above the clique x would correspond to the hidden parameters θ, while the graph would correspond to the observations y.

The problem is of course that we can't efficiently sample from this posterior distribution on cliques. In fact, there is not much unique about this problem- it is very often the case that posterior distributions are computationally hard to sample from. For this reason, Bayesian analysts often use efficient *approximations* for the posterior. For example, one might sample x from a random walk that converges to $p(x|G)$ in an exponential number of steps, but stop it much before convergence happens, or one might use some other type of algorithm to obtain some approximate distributions. Regardless, we can abstract this as sampling x by running $P(G)$ where $P(\cdot)$ is

some efficient probabilistic algorithm, and then estimating the utility of drug ℓ as $\mathbb{E}_{x \leftarrow P(G)} f_\ell(x)$. The problem is that since the clique problem is NP hard, with very high probability the output of $P(G)$ will *not* be a k clique and in fact, we can assume (using standard hardness of approximation results) that the set corresponding to x contains at least $k/2$ *non edges*. Using this observation, we can find a set S of about $2n^2/k$ *non edges* of G such that with probability close to 1 the following event E holds if we sample x from $P(G)$:

$$E = \{\exists (i,j) \in S \text{ s.t. } x_i x_j = 1\} \tag{2}$$

Now suppose that there is a drug that succeeds if and only if E occurs, then while we can plainly see that E will *never* happen for the actual posterior, the approximate Bayesian algorithm will think that the drug succeeds with probability close to 1! Note that this cannot be fixed by changing the notion of approximate distribution: every efficient probabilistic model will have the same issues!

The sos approach is different. Rather than producing an *actual distribution* that attempts to *approximate* the true posterior, we produce a *pseudo distribution* that *exactly matches* the low degree statistics of the posterior that we can easily derive from the data. In particular the sos pseudo-distribution will satisfy that $\tilde{\mathbb{E}}\, x_i x_j = 0$ for *every* non edge (i,j) over the graph as well as $\tilde{\mathbb{E}}(\sum x_i - k)^2 = 0$, even though the only actual distributions over $\{0,1\}^n$ that satisfy this condition are supported over k cliques. If the graph has a unique k clique x^*, then the only true distribution supported on k cliques satisfies $\mathbb{E}\, x_i = x_i^*$ for every i and so in particular $\mathbb{E}\, x_i$ equals either zero or one. Generally speaking, initially the sos pseudo-distribution will *not* satisfy this property, as it will reflect the *uncertainty* that a computationally bounded observer has about the clique, but as we increase the degree parameter d, for every i the value of $\mathbb{E}\, x_i$ will become closer and closer to either zero or one, until eventually it will converge to x_i^*, see the figure below:

For every function f, if our observations imply using a low degree sos argument that $f(x) \leq \alpha$, then our estimate $\tilde{\mathbb{E}}_x f(x)$ will be at most α. This yields a strategy that, in analogous way to the optimal Bayesian posterior, can't be "easily dutch booked" in the sense that one could not find f, g such that there is a simple proof that $f \leq g$ but $\tilde{\mathbb{E}}\, f > \tilde{\mathbb{E}}\, g$ where "simple" is defined as low degree sos. (See also our discussion on economics below.) As a computationally bounded strategy, sos can and will be wrong in its predictions, but at least its consistently wrong within a well defined system of making inferences, and provides some non trivial rigorous guarantees about its quality.

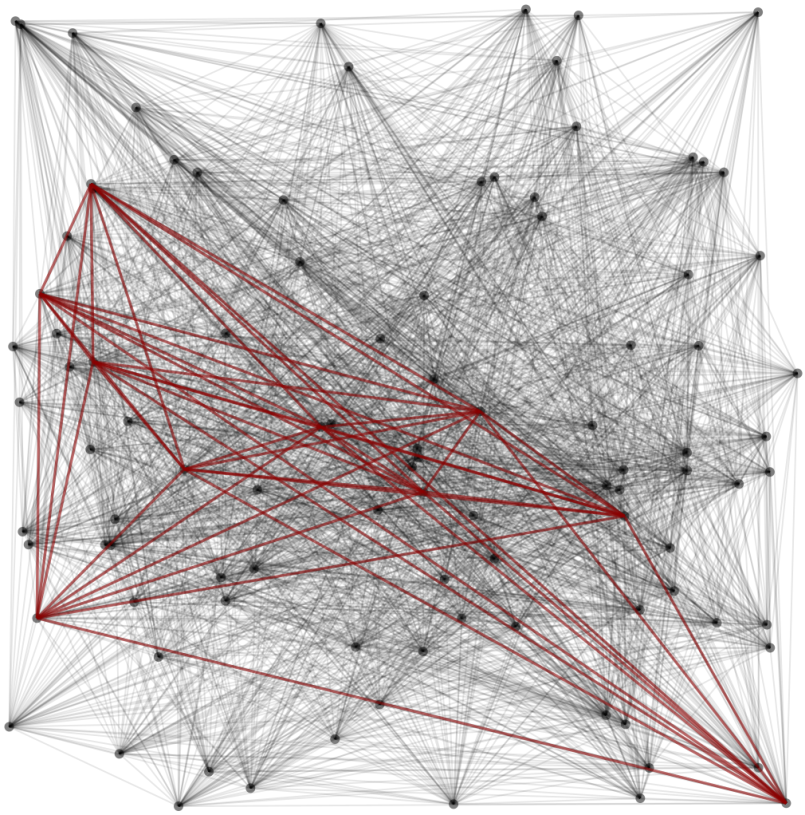

Figure 1: A random graph with a hidden clique. The sum-of-squares algorithm maintains a set of beliefs about which vertices belong to the hidden clique. Despite learning no new information, as we invest more computation time, the algorithm reduces uncertainty in the beliefs by making them consistent with increasingly powerful proof systems. Initially the beliefs have maximum uncertainty and correspond to the uniform distribution but they eventually converge on the correct hidden clique (red edges).

The economic view of Bayesian probabilities

There is a deep relation between Bayesian beliefs and *rational* strategy. The standard way to model an economic agent is as a computationally unbounded entity whose goal is to maximize its expected utility. This expectation is taken over the beliefs of the agent, which are based on combining a prior with the observations. Computational issues arise in economic settings as well.

One natural such setting is in the context of *prediction markets* (or stock markets more generally). Suppose there is a marketplace to place bets on some events $A_1, \ldots, A_n$ (for example, event A_i could be that candidate X wins in the i-th state in an election, or that company i outperforms its expectations) where each bet will give you one dollar if the corresponding event materializes. The question of setting the right price for each of these events is highly non trivial when they are *correlated*. For example, if we know that A_2 is implied by A_1 then we must set a price for A_1 higher than that for A_2 as otherwise we create an *arbitrage* opportunity (also known as *dutch book*) to make guaranteed profit without a risk. Indeed, the non-existence of such opportunities is a cornerstone of economics known as the Efficient Market Hypothesis; this can be proven if all participants are unbiased rational agents but the extent to which it applies in the real world is a subject of intense debate. However, what if an implication such as $A_1 \Rightarrow A_2$ is computationally hard to verify? For example, it may be that both A_1 and A_2 have some complex dependency structure on some underlying assets, in a way that the statement $A_1 \Rightarrow A_2$ would end up being equivalent to some SAT formula being unsatisfiable. In such a case we cannot expect the prices set by the market to always respect this implication.

SOS pseudo-distributions can be thought of as ensuring a "no computationally easy dutch book strategy". For example, think of a stock market that is based on some underlying set of events captured by $x \in \{0,1\}^n$, and where you can buy for every function $f : \{0,1\}^n \to \mathbb{R}$, a stock that would pay you $f(x)$ dollars when x is revealed. An arbitrage or dutch book strategy could arise when there are two functions such that $f(x) \geq g(x)$ for every x but the price of the stock corresponding to f is cheaper than the stock correpsonding to g. While in general we cannot be guaranteed that there is no such arbitrage, if we assign the price for the "f-stock" to be $\tilde{\mathbb{E}}\, f$ then at least we know that if the price of the f-stock is cheaper than the price of the g-stock then there is no short sos proof that $f \geq g$.

The existence of arbitrages in markets is an actual issue and there

are documented instances of arbitrage opportunities in prediction markets and stock markets (and there are companies that are spending immense computational resources in finding them and quickly exploiting them). In fact, finding and removing arbitrage opportunities is the main algorithmic challenge in creating combinatorial prediction markets that allow participants to place bets on arbitrarily complex combinations of underlying basic events, see also Kroer et al. [2016].

A recent paper Garrabrant et al. [2016] proposes a more general, Turing machine based, way of defining probabilities/prices for a "prediction market for mathematical statements" in a way to ensure that there is no computationally easy arbitrage/dutch book strategy regardless of the algorithm used to obtain it. Trying to compare the two approaches, as well as using other algorithmic frameworks as a source for defining computationally efficient Bayesian probabilities, is a very interesting research direction.

Other algorithms as a source for Bayesian probabilities.

It is an interesting and yet largely unexplored area to understand to what extent algorithms other than the sos algorithm give rise to internally consistent "probabilities" that can be interpreted as describing the beliefs of computationally bounded agents. When we run an algorithm for the task of recovering an unknown set of parameters x from observations y, even if the algorithm is not successful, it might be possible to interpret its *internal state* as codifying some *partial information* about x. Some algorithms, such as *belief propagation* come with fairly explicit mapping between the internal state and Bayesian probabilities. Similarly algorithms based on Monte Carlo Markov Chains or statistical-physics inspired algorithms that converge as a certain "temperture parameter" decreases can be interpreted as getting closer and closer to the ground truth. Recently, people have been able to interpret the internal state (i.e. intermediate neurons) of deep neural networks as encoding certain "beliefs" about the observed data (e.g., see Yosinski et al. [2015]).

Computational Bayesian probabilities and cryptography.

People sometimes describe a cryptographic scheme as offering, say, "128 bits of security", which roughly speaking means that one could not break the scheme using much fewer than 2^{128} computational

operations or with probability much better than 2^{-128}. This does not necessarily mean that the secret key is 128 bits long, though it definitely needs to be at least as long as that; for example, the security level of 1024 bit RSA encryption system is typically estimated at about 80 bits (Barker et al. [2007]).

Cryptography is typically most interested in "zero one scenarios" where one basically learns nothing about the underlying secrets using much less than 2^{128} operations and everything after spending more than that. The notion of pseudodistributions allows us to talk about "softer" scenarios where one can learn non-trivial information about the unknown data.

Beyond philosophy: the Bayes-Marley approach for analyzing SOS

How do we actually use this Bayesian perspective when we design algorithms or lower bounds? The intuition behind this is the following: while SOS is not coherent (i.e., does not correspond to actual distributions), it should be hard for any bounded observer to "catch" it being non coherent (at least up to some small error). So, when we design SOS based algorithm, we can pretend that these pseudo-distributions are actual distributions, as long as we limit ourselves to "bounded reasoning". Technically "bounded reasoning" corresponds to SOS proofs of low degree, but in practice it seems that most mathematical arguments we use in these contexts (except for the important exception of non constructive techniques such as the *probabilistic method*) can be "SOS'ed" with low degree. So, the paradigm for designing algorithms is:

> *Pretend you are working with actual distributions, avoid using the probabilistic method, and every little thing is going to be alright*

When designing *lower bounds* we need to construct pseudo distributions. These are not actual distributions, but it should not be easy for bounded observers to "catch" us in a discrepancy. So we need to respect any correlations that would be possible to bounded observers. That is, our approach is

> *If **you** can prove that the moments of a true posterior would have to satisfy property P, you'd better ensure that your pseudo distribution satisfies P as well.*

The nice thing about this that often the properties that a bounded observers can ascertain essentially fix the choice of the pseudo-distribution, and hence all that is left is the (often highly non trivial)

task of proving that this choice satisfies the pseudo-distribution constraints.

At the moment this might seem somewhat vague and abstract, but we will see actual examples of both upper and lower bounds using this approach.

References

Elaine Barker, William Barker, William Burr, William Polk, and Miles Smid. Nist special publication 800-57. *NIST Special Publication*, 800 (57):1–142, 2007.

Vincent Conitzer and Tuomas Sandholm. Computational criticisms of the revelation principle. In *Proceedings 5th ACM Conference on Electronic Commerce (EC-2004), New York, NY, USA, May 17-20, 2004*, pages 262–263, 2004. doi: 10.1145/988772.988824. URL http://doi.acm.org/10.1145/988772.988824.

William Feller. *An introduction to probability theory and its applications: volume I*, volume 3. John Wiley & Sons London-New York-Sydney-Toronto, 1968.

Scott Garrabrant, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. Logical induction. *arXiv preprint arXiv:1609.03543*, 2016.

Edwin T Jaynes. *Probability theory: The logic of science*. Cambridge university press, 2003.

Christian Kroer, Miroslav Dudík, Sébastien Lahaie, and Sivaraman Balakrishnan. Arbitrage-free combinatorial market making via integer programming. *arXiv preprint arXiv:1606.02825*, 2016.

Jason Yosinski, Jeff Clune, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. In *In ICML Workshop on Deep Learning*, 2015. arXiv preprint arXiv:1506.06579.