In many argument systems, a valid witness is defined as a polynomial f such that $C(f(x))$ vanishes on some set H, where C is some constraint. (As an intuitive example, if we wanted to enforce that $f\restriction_H$ contained only binary values, we could set $C(x) = x (x - 1)$.) Equivalently, f is a valid witness iff $C(f(x)) / Z_H(x)$ is a polynomial, where Z_H is the polynomial (of minimum degree) that vanishes on H.

If we’re using FRI, we could check the constraint by having the prover send (a low-degree extension of) f, then applying FRI to this quotient. Given a polynomial f that does not satisfy the instance, $C(f(x)) / Z_H(x)$ will not be a polynomial, but it may still have up to $\deg(C) \deg(f)$ points in common with a low-degree polynomial. To account for this, I think we would need $\delta \le 1 - \deg(C) \rho$ (δ is FRI’s proximity parameter, ρ is the code rate), which could make FRI rather expensive if C is high-degree.

Alternatively, we could have the prover send this quotient polynomial $q(x) = C(f(x)) / Z_H(x)$, have the verifier sample a random point r, “open” f and q at r, and check that $C(f(r)) = Z_H(r) q(r)$. Then by the Schwartz-Zippel lemma, invalid witnesses would be detected except with probability $\deg(C) \deg(f) / |\mathbb{F}|$. If we use a δ within the decoding radius, $(1 - \rho)/2$, then the Merkle roots of f and q can be treated as binding commitments to their (unique) proximate polynomials, $\tilde{f}$ and $\tilde{q}$, and we can verify openings of these proximate polynomials by running FRI on

$$\left{ \frac{f(x) - f(r)}{x - r}, \frac{q(x) - q(r)}{x - r} \right},$$

or sample a random α and run FRI on

$\frac{f(x) - f(r) + \alpha (q(x) - q(r))}{x - r}$.

If δ is outside the decoding radius, then we lose the binding property, but we can still use FRI as a sort of “weakly binding” polynomial commitment, since there are upper bounds on the number of polynomials within δ of any f or q. Let L such an upper bound (e.g. from the Johnson bound, or the conjectured bound in the DEEP-FRI paper). Then in a sense, the Merkle root of f binds the prover to (at most) L proximate polynomials, and likewise for q.

So we can apply FRI in the same way as before, and argue that $C(\tilde{f}(r)) = Z_H(r) \tilde{q}(r)$ is unlikely hold for any of the L^2 possible $(\tilde{f}, \tilde{q})$ pairs (if none of them are valid witnesses). However, this multiplies our soundness error by L^2, at least with a naive analysis (I think there may be ways to tighten this). In practice, most IOPs involve more than two prover polynomials, which would further increase the exponent on L.

So this error may greatly exceed our security parameter. I guess the obvious solution is to reduce soundness error by checking the polynomial identity at a bunch of points. We’re thinking about a recursive IOP with a small (64 bit) field and many (100+) witness polynomials, though, so the interpolations needed to open many points could get expensive.

One thing we could do is combine a set of 2^l polynomials (from the same prover message) into one higher-degree polynomial, which would increase L a bit but decrease its exponent. Constraints over the small polynomials would be compiled to constraints over the merged polynomial, with $f_i(g^j)$ mapped to $f_\mathrm{merged}(g^{2^l j + i})$. This seems a bit complex though, and it could end up being less efficient since it could increase the number of FRI rounds.

Are there better solutions? How do STARK implementations (or other FRI-based IOPs) handle this?