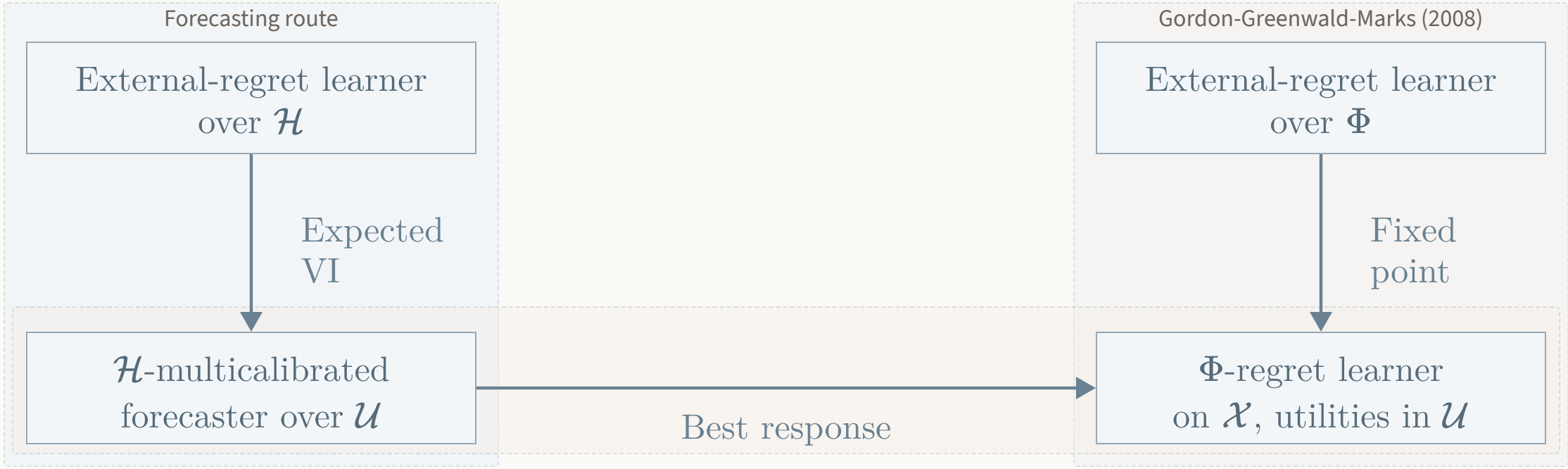
Forecasting and Calibration

Gabriele Farina

MIT 6.7980 · Topics in Multiagent Learning · Fall 2026

The important picture for today



Legend: Each arrow turns the algorithm in its starting box into the algorithm in its ending box using the optimization primitive written on the arrow.

Forecasting: What does a 70% chance of rain mean?

A forecaster reports p_t before seeing the weather: $y_t = 1$ means rain, and $y_t = 0$ means no rain.

Collect the days with a 70% forecast

It should rain on about 70% of those days.

Repeat this check for every probability the forecaster uses.

The above condition is called **calibration**.

How calibration became an online learning problem

> **1977 · Murphy & Winkler**
>
> study whether weather forecasters' probabilities agree with observed frequencies.

> **1982 · Dawid**
>
> develops calibration as a criterion for probability forecasts.

> **1985 · Oakes and Dawid**
>
> show that every deterministic forecasting rule fails calibration on some outcome sequence.

> **1998 · Foster & Vohra**
>
> construct randomized forecasts that are calibrated on every outcome sequence.

Murphy & Winkler (1977) · Dawid (1982) · Oakes (1985) · Foster & Vohra (1998).

More formally: Expected calibration error (ECE)

Consider a finite grid $\mathcal{Q}$ of forecasts, and $I_q = \{t : p_t = q\}, N_q = |I_q|, \bar{y}_q = \frac{1}{N_q} \sum_{t \in I_q} y_t$.

$$\begin{aligned} \mathrm{ECE}_T &:= \sum_{q \in \mathcal{Q}} \frac{N_q}{T} |\bar{y}_q - q| \\ &= \frac{1}{T} \sum_{q \in \mathcal{Q}} \left| \sum_{t \in I_q} (y_t - q) \right|. \end{aligned}$$

Absolute values are important

Overprediction in one forecast bucket cannot cancel underprediction in another.

Impossibility when Nature can use the forecast adversarially

It is impossible to control ECE against an adversarial process. After seeing p_t, Nature makes it rain when $p_t < 1/2$ and keeps the day dry otherwise:

$$y_t = \begin{cases} 1, & p_t < 1/2, \\ 0, & p_t \geq 1/2. \end{cases}$$

Every forecast value that gets used is wrong

Forecasts below $1/2$ underestimate rain by more than $1/2$. Forecasts at least $1/2$ overestimate it by at least $1/2$.

*What changes if Nature sees the forecast **distribution**, but not the forecast drawn from it?*

Oakes (1985), Dawid (1985); Foster & Vohra (1998), introduction.

Calibration without a weather model

Foster and Vohra's result

Using past outcomes and private randomization, a forecaster can make its long-run calibration error arbitrarily small on every outcome sequence.

The algorithm needs no atmospheric measurements or meteorological model. It also works when the process generating the outcomes changes over time.

Is this a positive or a negative result?

Foster & Vohra (1998), Asymptotic Calibration.

Intuition: a forecaster can “cook the books”

Simplified example: imagine that the calibration error was measured as

$$\frac{1}{T}\sum_{t=1}^{T}(y_t - p_t)$$

There exists a simple way of ensuring the above quantity $\to 0$. *How?*

Multicalibration protocol

1. Observe $c_t \in \mathcal{C}$, the information available before forecasting.
2. Announce a distribution D_t over forecasts $p \in \mathcal{U}$.
3. Nature chooses the outcome $u_t \in \mathcal{U}$.
4. Observe u_t and update the learner.

Keep the sampled forecast private

Nature sees D_t, but must choose u_t before seeing the sampled forecast. The expectations below average over this draw.

Both forecasts and outcomes lie in a compact convex set $\mathcal{U} \subseteq \mathbb{R}^d$. We assume the tests are bounded and measurable.

Write the error for a general test

A test $h : \mathcal{C} \times \mathcal{U} \to \mathbb{R}^d$ measures the error $u_t - p$. It gives a weight for scalar errors and a direction for vector errors:

$$\mathrm{MC}_T(h) := \sum_{t=1}^{T} \mathbb{E}_{p \sim D_t} \left[\langle h(c_t, p), u_t - p \rangle \right].$$

We want

$$\sup_{h \in \mathcal{H}} \mathrm{MC}_T(h) = o(T).$$

If $\mathcal{H}$ contains $-h$ whenever it contains h, the bound also controls $|\mathrm{MC}_T(h)|$. We can include the zero test.

ECE is a special case of multicalibration

Ignore the context c and choose a separate sign for each forecast value:

$$h_s(c, p) := \sum_{q \in \mathcal{Q}} s_q \mathbf{1}\{p = q\}, \qquad s \in \{-1, 1\}^{\mathcal{Q}}.$$

For the observed forecasts,

$$\frac{1}{T} \max_s \sum_{t=1}^{T} h_s(c_t, p_t)(y_t - p_t) = \sum_{q \in \mathcal{Q}} \frac{N_q}{T} |\bar{y}_q - q| = \mathrm{ECE}_T \,.$$

Choose each sign to make that bucket's accumulated error nonnegative.

From forecasts to decisions

As it turns out, multicalibration has a flavor similar to Φ-regret minimization. We will see how both can be reduced to

(regret minimization) + (a nonlinear optimization primitive)

Recall Φ-regret

Choose x in a compact convex set $\mathcal{X} \subseteq \mathbb{R}^d$ and earn $\langle x, u_t \rangle$. Compare this with playing $\phi(x)$, for each fixed deviation $\phi \in \Phi$.

$$R_T^\Phi(\phi) := \sum_{t=1}^{T} \mathbb{E}_{x \sim \mu_t} [\langle \phi(x) - x, u_t \rangle], \qquad \phi \in \Phi.$$

External regret

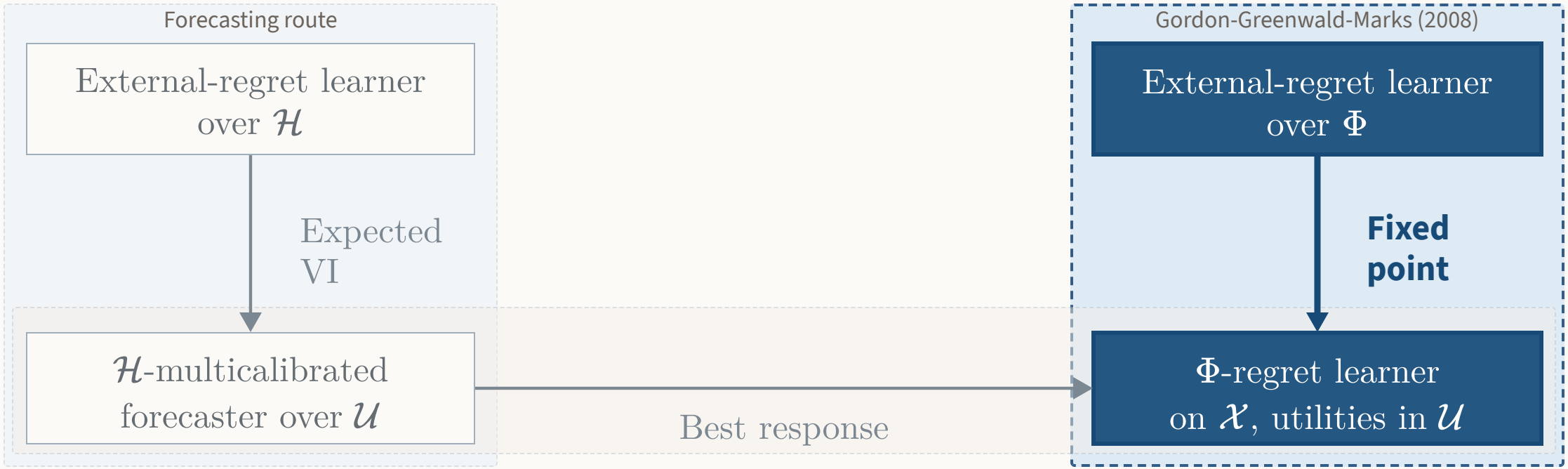
$\phi(x) = \hat{x}$ for a fixed action $\hat{x}$.

Swap regret

On a simplex, $\phi(x) = Qx$ for a column-stochastic matrix Q.

We want $\sup_{\phi \in \Phi} R_T^\Phi(\phi) = o(T)$.

First route: Gordon-Greenwald-Marks



Ask the learner for ϕ_t, play $x_t = \phi_t(x_t)$, then return the utility

$$G_t(\phi) := \langle \phi(x_t), u_t \rangle.$$

Gordon, Greenwald & Marks (2008). This version uses exact fixed points. We assume an efficient fixed-point oracle.

The fixed point makes the two regrets equal

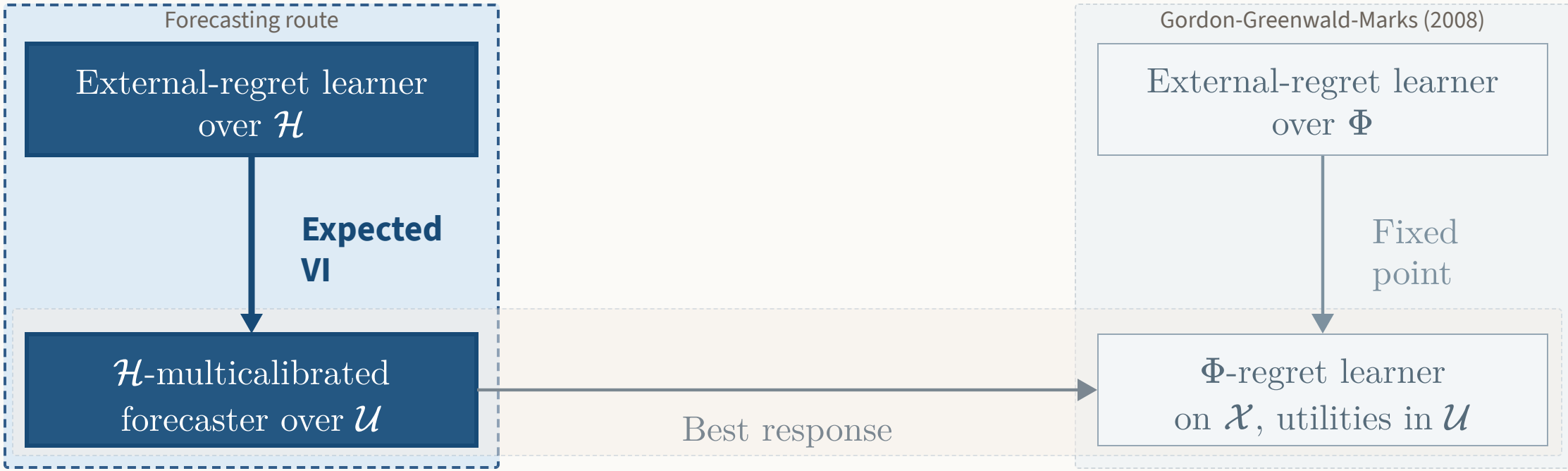


Fix any comparator $\phi \in \Phi$. The external regret against it is

$$\begin{aligned}\mathrm{Reg}_{\Phi,T}(\phi) &= \sum_t \big[G_t(\phi) - G_t(\phi_t)\big] \\ &= \sum_t \langle \phi(x_t) - \phi_t(x_t), u_t \rangle.\end{aligned}$$

$$\phi_t(x_t) = x_t \quad \Longrightarrow \quad \boxed{\mathrm{Reg}_{\Phi,T}(\phi) = R_T^{\Phi}(\phi).}$$

Learn calibrated forecasts



Choose forecasts that control the current test

The no-regret learner suggests which test to pay attention to. We then try to "neutralize" it.

Expected variational inequalities

Given $S : \mathcal{U} \to \mathbb{R}^d$, find a distribution D over $\mathcal{U}$ satisfying

$$\boxed{\mathbb{E}_{p\sim D}[\langle S(p), u - p\rangle] \leq \epsilon \qquad \forall u \in \mathcal{U}.}$$

The bound must hold for every u

The operator S may be discontinuous. A distribution over forecasts can satisfy the bound.

We assume an EVI solver. Its cost and accuracy depend on the domain and how we can access S.

An algorithm for multicalibration

Run an external-regret learner on $\mathcal{H}$ (or its convex hull).

1. Observe c_t and ask the learner for h_t.
2. Solve an ϵ_t-EVI for $S_t(p) = h_t(c_t, p)$; output D_t.
3. Observe u_t and send the linear utility below to the learner.

$$g_t(h) := \mathbb{E}_{p \sim D_t}\left[\langle h(c_t, p), u_t - p \rangle\right].$$

The EVI bounds the score of today's test

$$g_t(h_t) \leq \epsilon_t.$$

The EVI inequality applies to the realized u_t.

Farina & Perdomo (2026).

Calibration is bounded by external regret



Fix $h \in \mathcal{H}$. Add and subtract the learner's scores:

$$\mathrm{MC}_T(h) = \underbrace{\sum_t \left[g_t(h) - g_t(h_t)\right]}_{\mathrm{Reg}_{\mathcal{H},T}(h)} + \sum_t g_t(h_t)$$

$$\leq \mathrm{Reg}_{\mathcal{H},T}(h) + \sum_t \epsilon_t.$$

Apply $g_t(h_t) \leq \epsilon_t$ in each round.

From multicalibration to Φ-regret



If the prediction is good...

... we might as well best respond to the prediction.

Best respond to the forecast

Forecast the next utility vector $p \in \mathcal{U}$. Choose a best response:

$$\sigma(p) \in \arg\max_{x \in \mathcal{X}} \langle x, p \rangle.$$

Draw $p \sim D_t$ $\longrightarrow$ Play $x = \sigma(p)$ $\longrightarrow$ Observe u_t

The forecasts determine the action distribution

Let μ_t be the distribution of $\sigma(p)$ for $p \sim D_t$. Use a fixed rule to break ties.

Announce μ_t before Nature chooses u_t, and keep the sampled action hidden. Taking the mean action can change regret for nonlinear deviations.

Which multicalibration tests to use?

For each $\phi \in \Phi$, the test is the change it makes to our chosen action:

$$\boxed{h_\phi(p) := \phi(\sigma(p)) - \sigma(p).}$$

With context-dependent deviations, use $h_\phi(c, p) = \phi(c, \sigma(p)) - \sigma(p)$.

Bound Φ-regret by calibration error



$$R_T^{\Phi}(\phi) = \sum_t \mathbb{E}_{p\sim D_t}\langle h_\phi(p), u_t\rangle$$
$$= \underbrace{\sum_t \mathbb{E}\langle h_\phi(p), u_t - p\rangle}_{\mathrm{MC}_T(h_\phi)} + \underbrace{\sum_t \mathbb{E}\langle h_\phi(p), p\rangle}_{\leq 0}.$$

(Last step because $\langle h_\phi(p), p\rangle = \langle \phi(\sigma(p)), p\rangle - \langle \sigma(p), p\rangle \leq 0$.)

Since $\sigma(p)$ is a best response and $\phi(\sigma(p)) \in \mathcal{X}$, the forecast term is nonpositive.

Combine the two forecasting reductions



$$R_T^{\Phi}(\phi) \leq \mathrm{MC}_T(h_\phi) \leq \mathrm{Reg}_{\mathcal{H},T}(h_\phi) + \sum_t \epsilon_t.$$

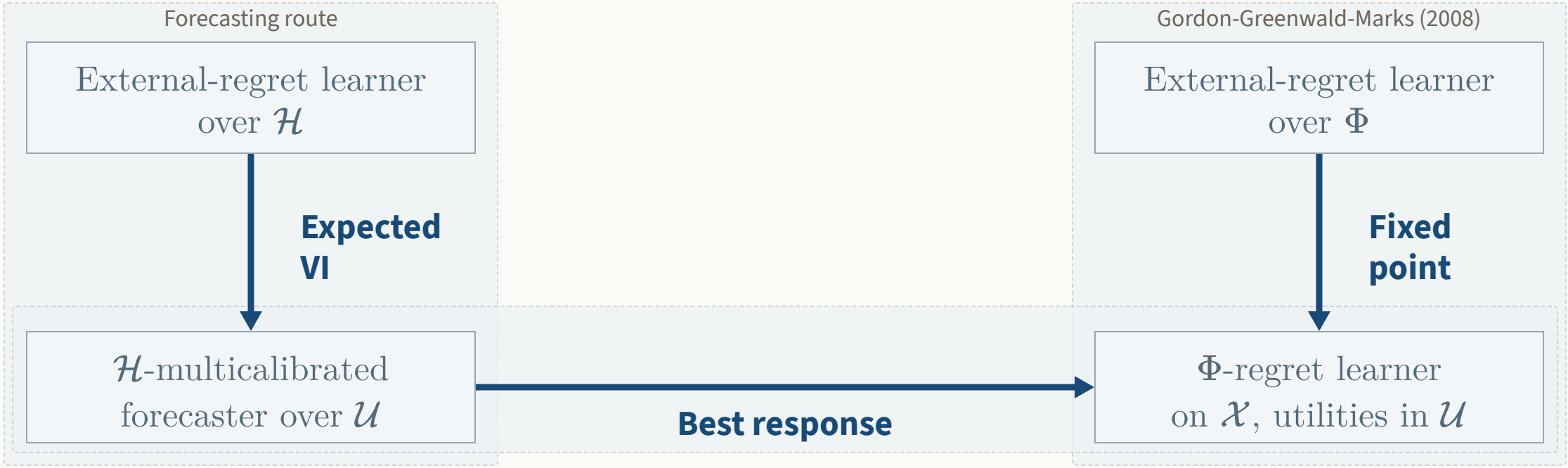
Two routes from external regret

Ingredient	Gordon-Greenwald-Marks	Forecasting route
Learn over	Valid deviations Φ	Tests $\mathcal{H} \supseteq \mathcal{H}_\Phi$
Optimization step	Fixed point of ϕ_t	EVI for h_t
Play	The fixed point x_t	Best response $\sigma(p)$
Regret bound	Exact identity	Upper bound + EVI error

Forecasting route is more permissive

No geometry of endomorphisms $X \to X$ needs to be taken into account.

Summary


Forecasting route
External-regret learner over $\mathcal{H}$
Expected VI
$\mathcal{H}$-multicalibrated forecaster over $\mathcal{U}$
Gordon-Greenwald-Marks (2008)
External-regret learner over Φ
Fixed point
Φ-regret learner on $\mathcal{X}$, utilities in $\mathcal{U}$
Best response