Model Selection for Estimating Propensity Scores

Greg Ridgeway

http://www.i-pensieri.com/gregr

RAND Statistics Group, Santa Monica, CA

Joint work with Dan McCaffrey and Andrew Morral

Propensity scores in a nutshell

- y_0 and y_1 are potential outcomes
- $t = 1$ indicates assignment to treatment, $t = 0$ indicates assignment to control
- $p_i = P(t = 1|\mathbf{x}_i)$ is the propensity score
- $w_i = p_i/(1 - p_i)$ propensity score weight
- The treatment effect on the treated:

$$\mathrm{E}(y_1 - y_0|t = 1) \approx \frac{\sum t_i y_{1i}}{\sum t_i} - \frac{\sum w_i(1-t_i)y_{0i}}{\sum w_i(1-t_i)}$$

Estimating propensity scores

- Let $\log \frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \beta_0 + \sum_{j=1}^{J} \beta_j h_j(\mathbf{x})$
- Consider the regularized logistic regression likelihood

$$\sum t_i \beta' \mathbf{h}(\mathbf{x}_i) - \log(1 + \exp(\beta' \mathbf{h}(\mathbf{x}_i))) - \lambda \sum_{j=1}^{J} |\beta_j|$$

  - $\lambda = 0$ yields the usual logistic regression
  - $\lambda = \infty$ yields $p(\mathbf{x}) = \bar{t}$, the baseline rate
  - For other λ, many of the optimal βs are 0

Implemention

- We let h_j be the collection of up to 3-way interactions of indicator functions involving $\mathbf{x}$. For example,

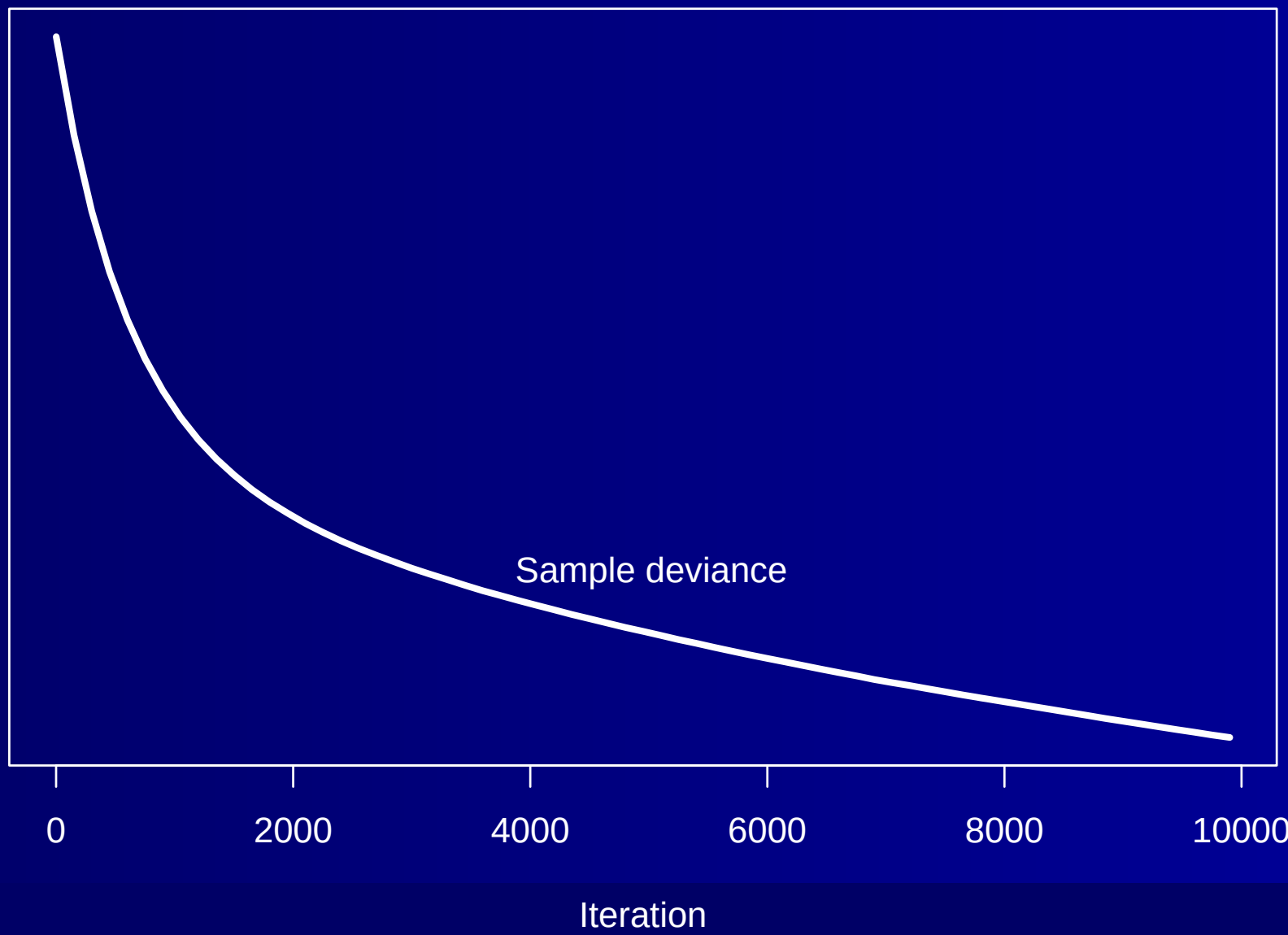
  $$h_1(\mathbf{x}) = I(\text{age} < 15)I(\text{SFI} > 10)I(\text{probation} = \text{yes})$$

- Forward stagewise regression:

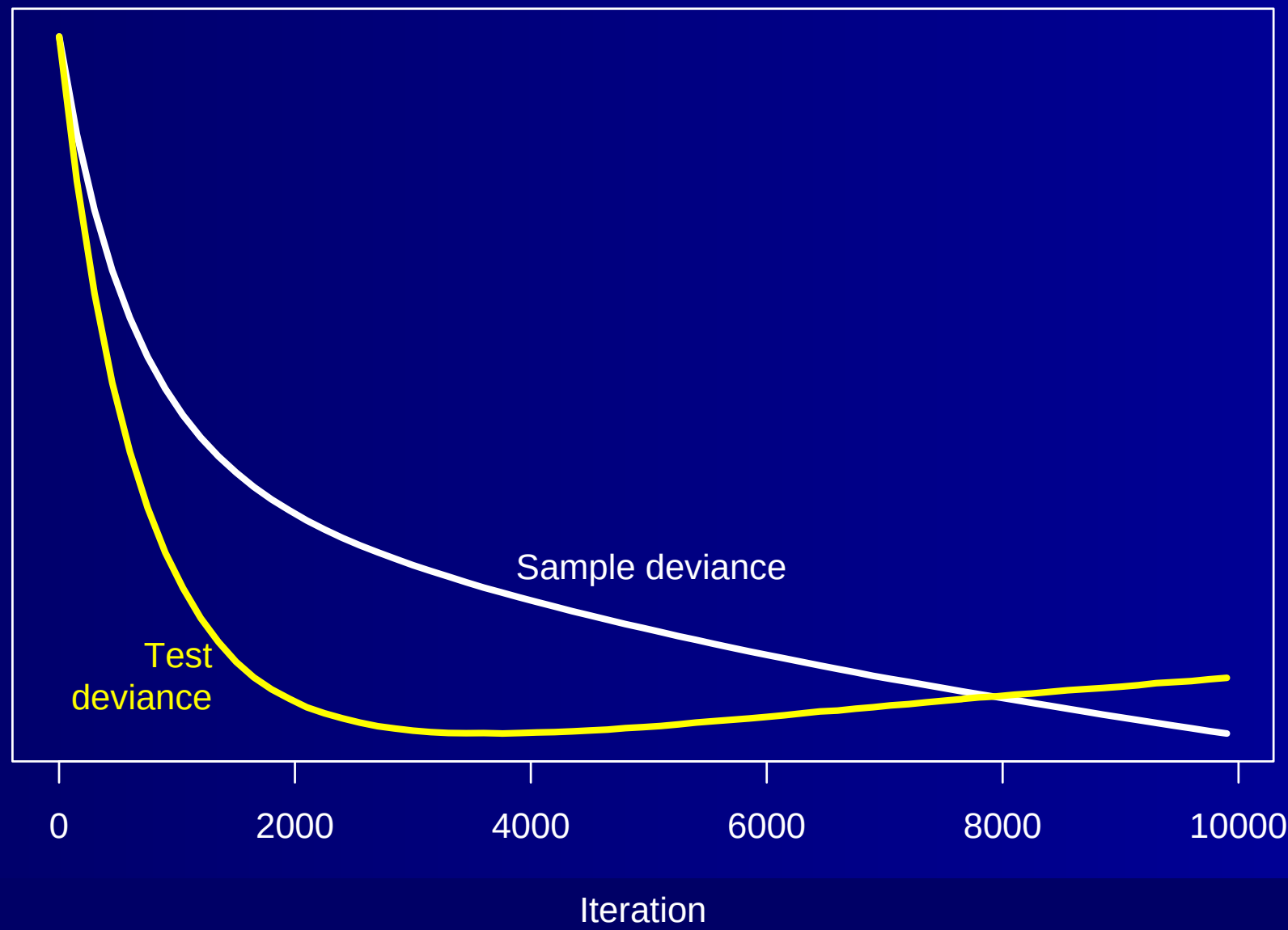
  $$\log \frac{p(\mathbf{x})}{1-p(\mathbf{x})} \leftarrow \log \frac{p(\mathbf{x})}{1-p(\mathbf{x})} + \epsilon h_*(\mathbf{x})$$

  where $\epsilon = 0.001$ and $h_*(\mathbf{x})$ is the basis function most correlated with the residuals of the current fit

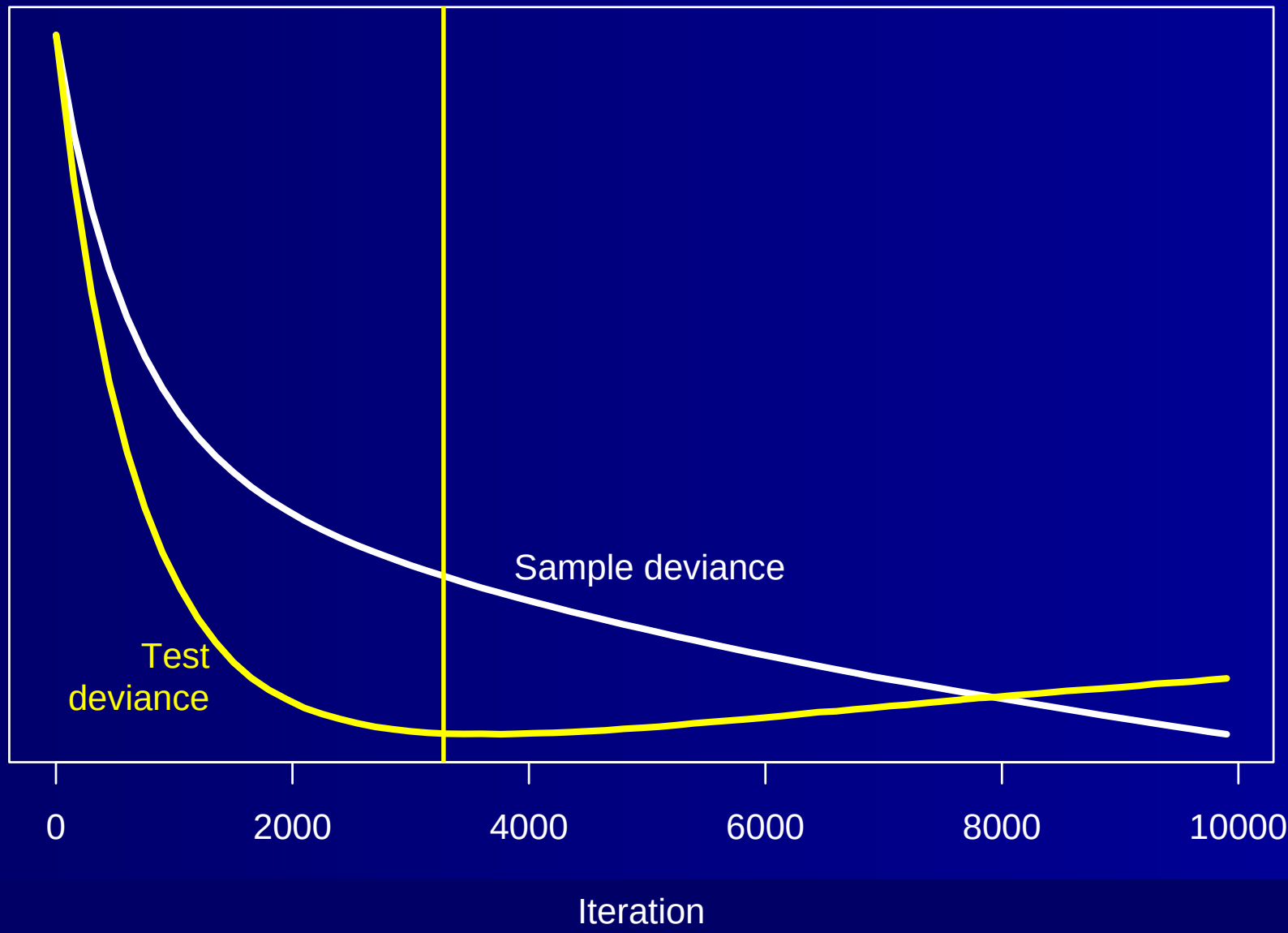
Each iteration decreases deviance



$p(\mathbf{x})$ can be overfit



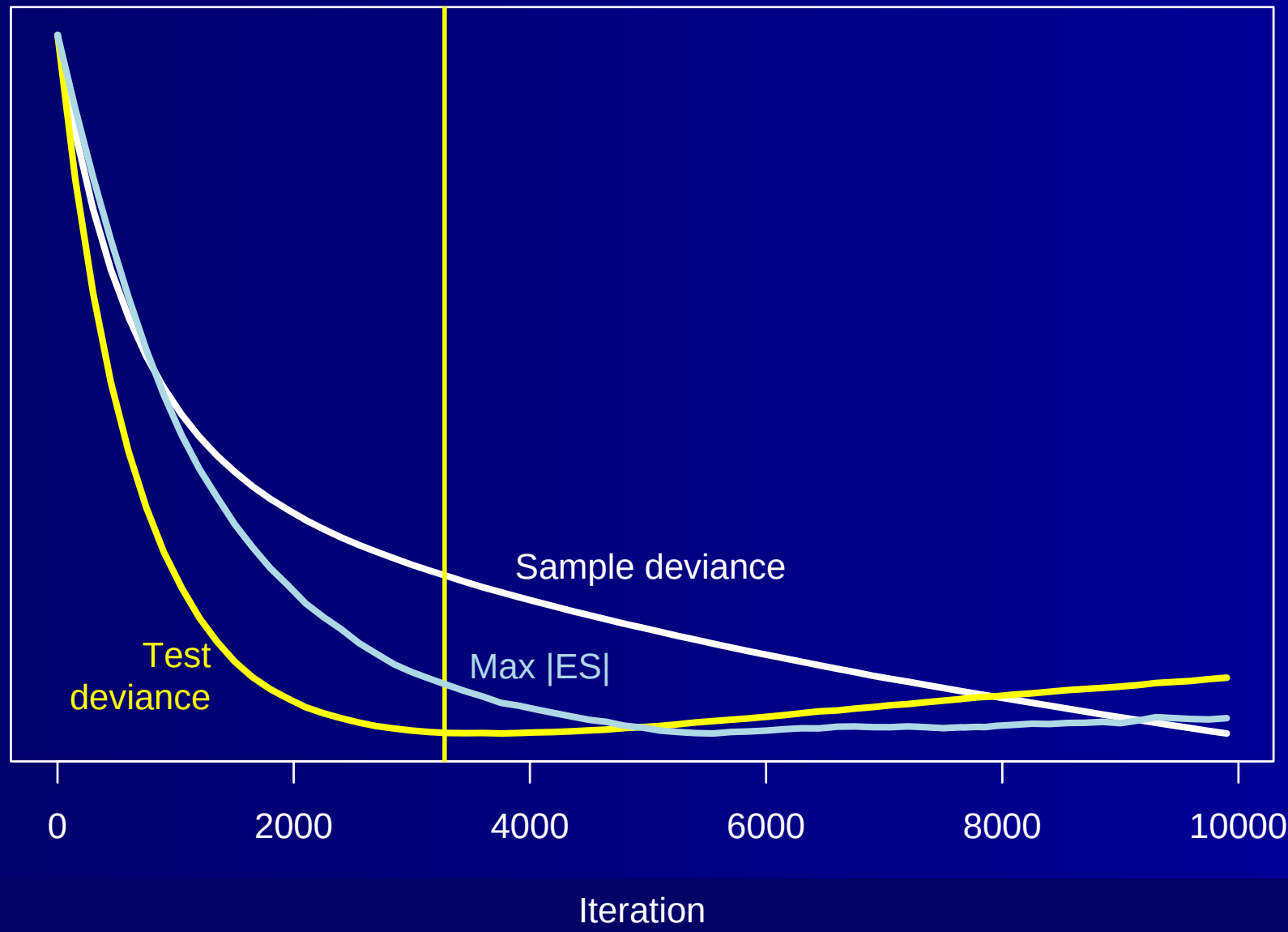
10-fold CV minimizes deviance



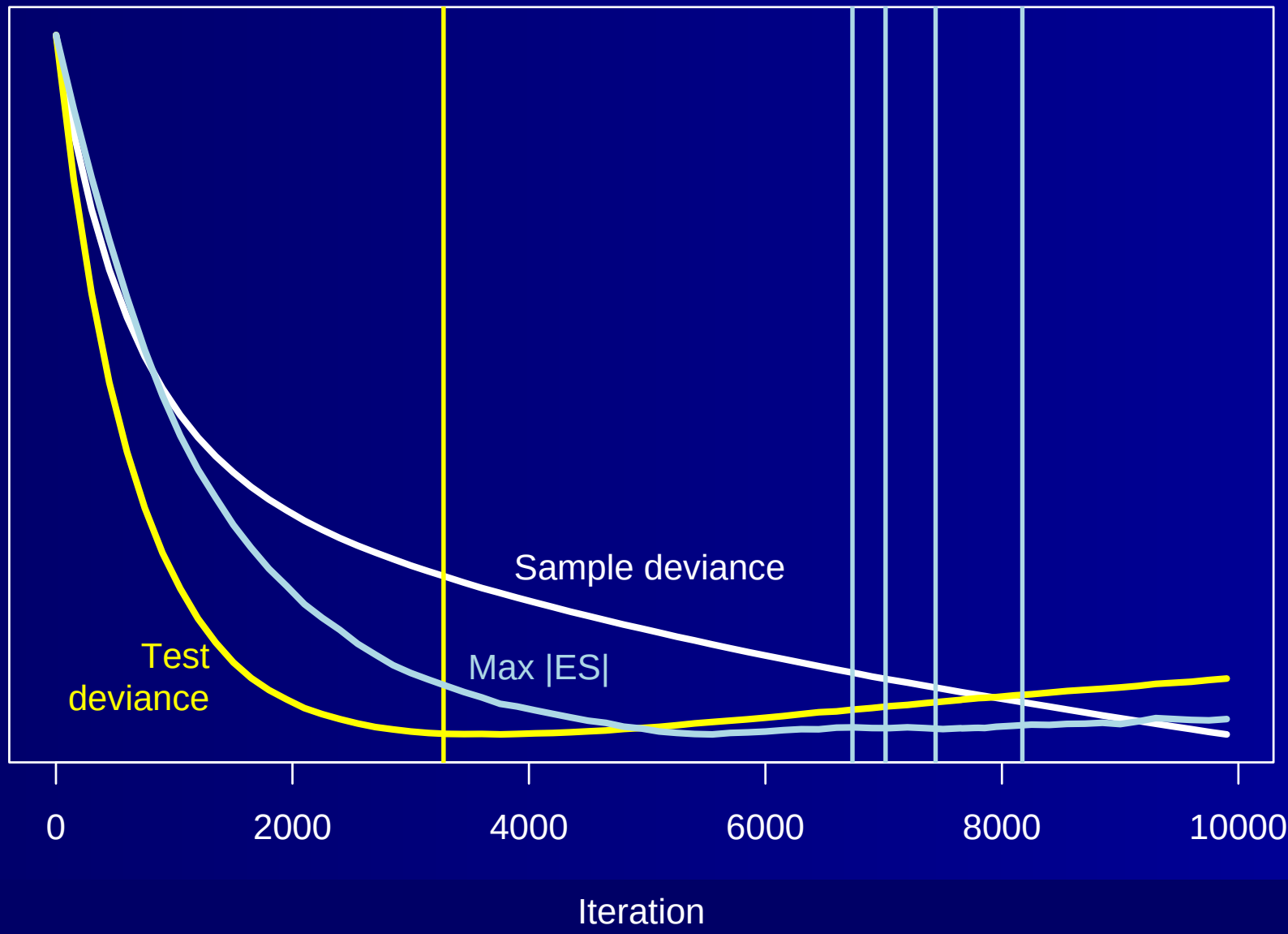
Effect size measure of balance

Variable	weighted treatment mean	weighted control mean	effect size weighted	effect size unweighted
Treatment motivation	2.52	2.22	0.23	0.89
Environmental risk	30.61	31.09	-0.05	0.17
Substance use	7.61	6.94	0.16	0.69
Complex behavior	12.84	13.00	-0.02	0.09
Age	15.82	15.76	0.07	0.56
⋮			⋮	⋮
Average |ES|			**0.107**	**0.307**
Max |ES|			**0.260**	**1.070**

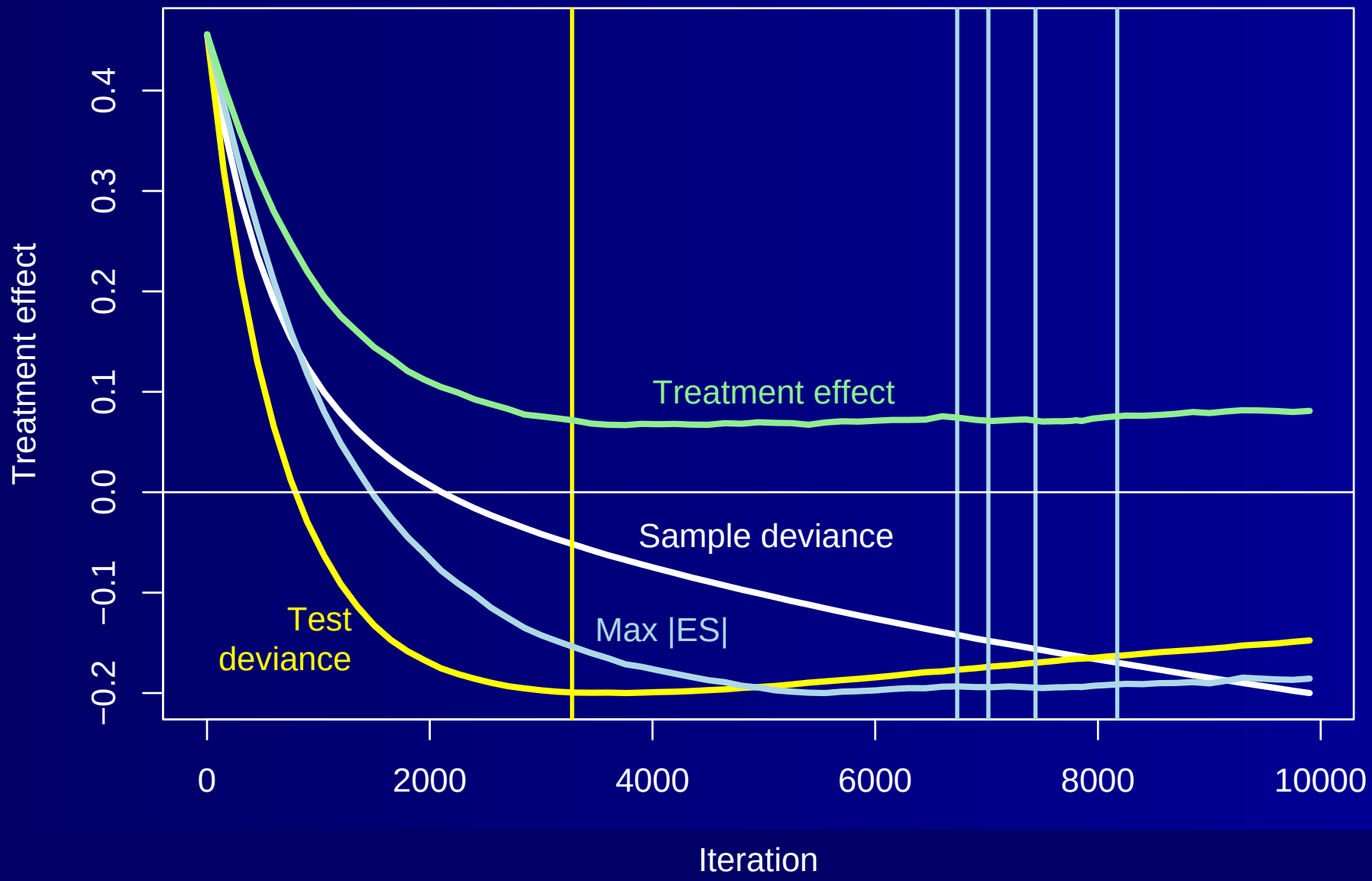
Balance encourages overfit



Similar for other stopping rules

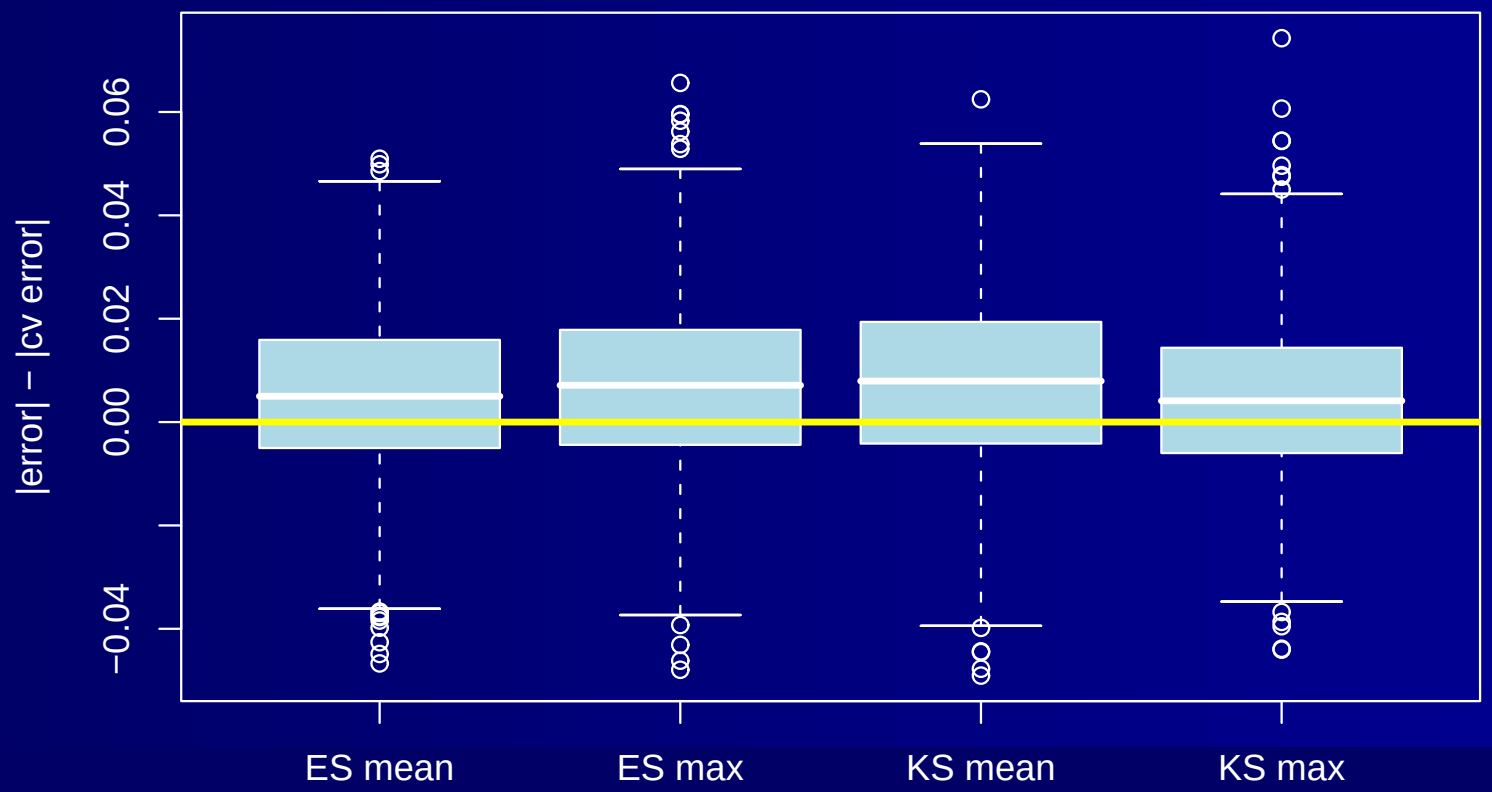


Compare with treatment effect

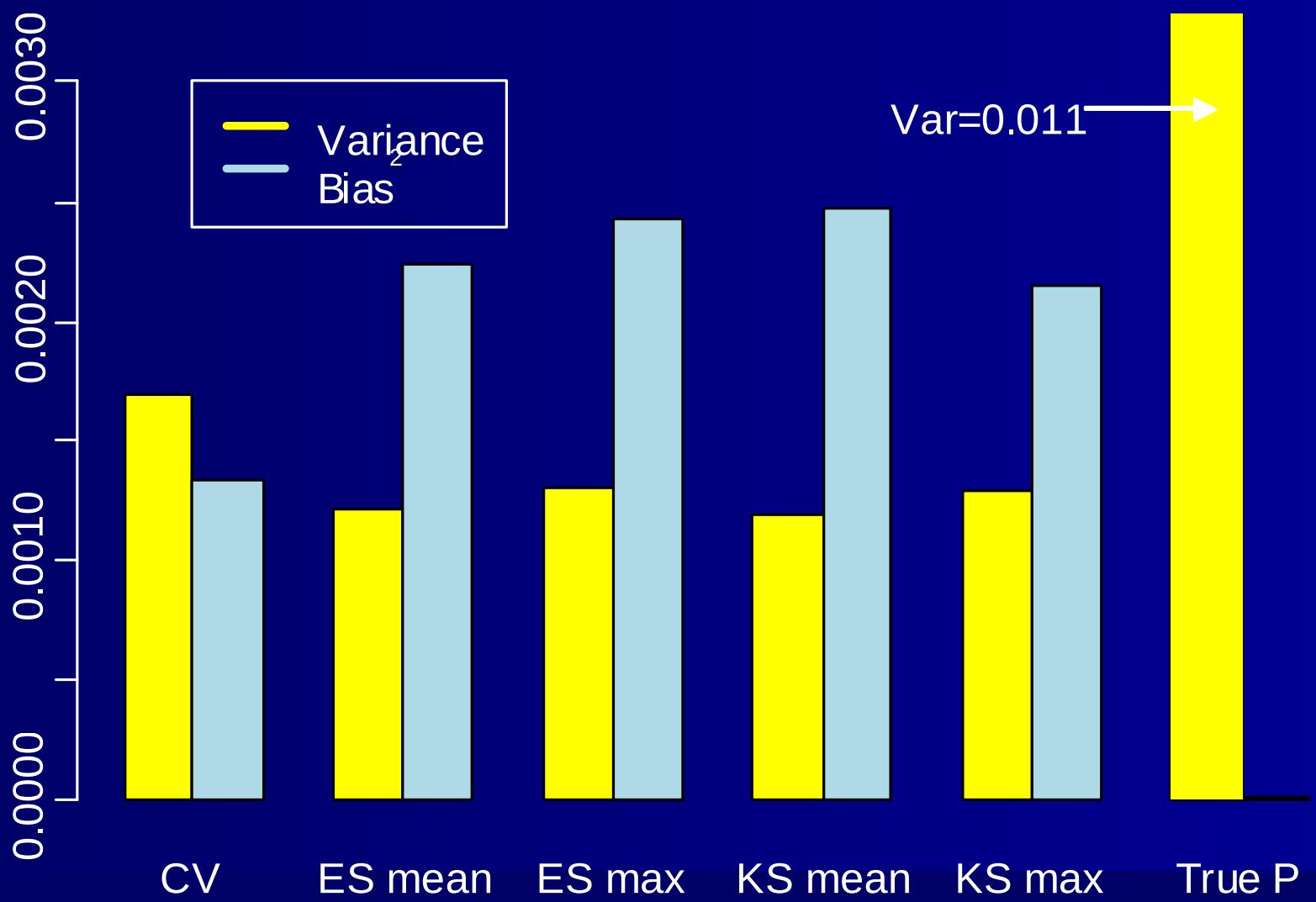


Performance of stopping rules

- Within replication absolute error is greater for non-likelihood based selection criteria
- 10-fold CV ranked #1 in 71% of replications



Bias variance tradeoff



Implications

- Stopping rules that measure fit are appealing for communicating the reduction of bias
- Simulation seems to indicate that effort in getting good propensity scores results in better estimates of treatment effects
- Competing factors: ratio bias, ANCOVA-style variance reduction, balance on irrelevant covariates