

Unvermeidliche Divergenz des Selektionsbias bei iterativer Suchraum-Eskalation

nick.albrecht

January 2026

1 Introduction

[a4paper,11pt]article

amsmath,amssymb,amsthm geometry hyperref microtype margin=2.5cm

Unvermeidliche Divergenz des Selektionsbias bei iterativer Suchraum-Eskalation

Abstract

Es wird gezeigt, dass in datengetriebenen Forschungs- und Optimierungsprozessen mit iterativer, adaptiver Suchraum-Erweiterung die effektive Anzahl getesteter Varianten K zwangsläufig divergiert. Da der erwartete Selektionsbias als Extremwert monoton in $\ln K$ wächst, folgt, dass der Bias ohne obere Schranke anwächst. In Kombination mit wiederholten $\arg \max$ -Operationen auf Zufallsresiduen impliziert dies, dass jede endliche Performancekennzahl mit hinreichender Wahrscheinlichkeit erreicht werden kann, selbst bei reiner Nullhypothese ohne echten Edge.

2 Grundannahmen (Nullmodell)

Es wird ein idealisiertes, analytisch tractables Setting betrachtet:

- Es existiert kein echter Edge:

$$\mu_{true} = 0.$$

- Jede getestete Variante k liefert eine geschätzte, standardisierte Performance

$$Z_k \sim \mathcal{N}(0, 1), \quad k = 1, 2, \dots$$

- Die Zufallsvariablen Z_k sind unabhängig.
- Der Forschungsprozess selektiert stets die beste beobachtete Variante:

$$Z_{\max}(K) := \max_{1 \leq k \leq K} Z_k.$$

Diese Annahmen sind konservativ und begünstigen eher eine Unterschätzung des Effekts.

3 Extremwertgrundlage

Für K i.i.d. standardnormalverteilte Zufallsvariablen gilt asymptotisch:

$$E[Z_{\max}(K)] \sim \sqrt{2 \ln K} \quad \text{für } K \rightarrow \infty.$$

Der Erwartungswert des selektierten Maximums ist damit streng monoton wachsend in K und unbeschränkt nach oben. Die Analyse des Selektionsbias reduziert sich folglich auf die Frage, ob der Forschungsprozess ein unbeschränktes Wachstum von K zulässt.

4 Iterative Suchraum-Eskalation

4.1 Modellierung iterativer Optimierung

Es wird ein iterativer Prozess $r = 1, 2, \dots$ betrachtet, in dem aufgrund unzureichender Zwischenergebnisse sukzessive neue methodische Freiheitsgrade eingeführt werden, etwa:

- neue Schwellenwerte,
- neue Features oder Diskretisierungen,
- neue Regeln oder Regelkombinationen,
- neue Filter oder Zielfunktionen.

Sei $c_r \geq 2$ die Anzahl neu unterscheidbarer Varianten, die in Iteration r hinzukommen. Die kumulative Suchraumgröße nach r Iterationen ist dann:

$$K_r = \prod_{i=1}^r c_i.$$

Falls $c_i \geq 2$ für unendlich viele Iterationen gilt, folgt:

$$K_r \geq 2^r, \quad \ln K_r \geq r \ln 2 [r \rightarrow \infty] \infty.$$

4.2 Bemerkung zur Methodik

Entscheidend ist, dass diese Erweiterungen datenabhängig erfolgen. Jede Anpassung wird durch beobachtete Residuen motiviert und ist damit Teil des effektiven Suchraums, auch wenn sie nicht explizit als zusätzlicher Test gezählt wird.

5 Divergenz des Selektionsbias

Durch Einsetzen des Wachstums von K_r in die Extremwertnäherung ergibt sich:

$$E[Z_{\max}(K_r)] \sim \sqrt{2 \ln K_r} \geq \sqrt{2r \ln 2} [r \rightarrow \infty] \infty.$$

[Divergenz des Selektionsbias] Bei iterativer, adaptiver Suchraum-Erweiterung mit unbeschränktem Wachstum von K divergiert der erwartete Selektionsbias gegen unendlich.

6 Erreichbarkeit beliebiger Performancelevel

Sei $M > 0$ eine beliebige endliche Performanceschwelle. Dann gilt:

$$P(Z_{\max}(K) \geq M) = 1 - \Phi(M)^K,$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet.

Für $K \rightarrow \infty$ folgt:

$$\Phi(M)^K \rightarrow 0, \quad \Rightarrow \quad P(Z_{\max}(K) \geq M) \rightarrow 1.$$

[Beliebige Performance ist erreichbar] Für jede endliche Performanceschwelle $M > 0$ gilt:

$$\lim_{K \rightarrow \infty} P(Z_{\max}(K) \geq M) = 1.$$

Damit ist es bei hinreichend großem Suchraum nahezu sicher, eine scheinbar beliebig gute Performance zu beobachten, selbst unter reiner Nullhypothese.

7 Rolle iterativer Optimierung

Eine bloße Erhöhung der Anzahl expliziter Versuche führt zu

$$K = n \quad \Rightarrow \quad \ln K = \ln n,$$

was lediglich ein langsames Bias-Wachstum impliziert.

Iterative Freiheitsgraderweiterung erzeugt hingegen typischerweise:

$$K \sim 2^r \quad \Rightarrow \quad \ln K = \Theta(r),$$

wodurch sich die Wachstumsordnung des Bias qualitativ verschiebt.

Schlechte Zwischenergebnisse verschärfen den Effekt zusätzlich, da sie weitere Iterationen und damit zusätzliche Freiheitsgrade auslösen. Der Selektionsbias wirkt ausschließlich auf das finale, konditionierte Ergebnis.

8 Hauptresultat

[Unvermeidliche Beliebigkeit] In einem datengetriebenen Prozess mit iterativer, adaptiver Suchraum-Erweiterung gilt:

1. $K \rightarrow \infty$,
2. $\ln K \rightarrow \infty$,
3. $E[Z_{\max}] \rightarrow \infty$,
4. jede endliche Performancekennzahl ist mit Wahrscheinlichkeit 1 erreichbar.

9 Schlussfolgerung

Ohne ex-ante Fixierung des Suchraums, harte Begrenzung iterativer Freiheitsgrade und strikt separierte Validierung ist jede beobachtete Performance prinzipiell vollständig kompatibel mit reinen Zufallsresiduen. Hohe Performannewerte stellen in diesem Regime keinen Nachweis für Information dar, sondern sind eine zwangsläufige Konsequenz unbeschränkter Suchraum-Eskalation.