

Interchangeability in Bayesian Security Games: Exact Balance and Generic Uniqueness

September 6, 2026

Abstract

We study interchangeability of Bayes–Nash equilibria in simple security games with type-independent defender payoffs and strictly positive coverage effects. Failure of interchangeability requires an exact balance between two disjoint groups of attacker types and two disjoint groups of targets. Consequently, for fixed defender payoff gains, interchangeability holds outside a finite union of hyperplanes in the prior simplex, uniformly over all strictly monotone attacker payoff tables. With one defender resource, the defender’s equilibrium strategy is generically unique. We also prove that failure requires at least three targets and three types. Matching three-target, three-type examples show failure even with pairwise distinct type probabilities, and nonconvexity of the defender equilibrium strategy set. For general catcher–evader games, a short proof of aggregate invariance yields one-sided interchangeability and recovers the earlier non-Bayesian multiple-resource theorem. An appendix identifies the precise limitation of its extension to heterogeneous types.

1 Introduction

Equilibria are *interchangeable* if combining either player’s strategy from one equilibrium with the other player’s strategy from another still yields an equilibrium. Korzhyk, Conitzer, and Parr proved this property for non-Bayesian security games with multiple attacker resources [1, Theorem 8]. Li, Conitzer, and Korzhyk introduced the broader catcher–evader framework and stated the same conclusion there [2, Theorem 3]. The latter conclusion requires qualification: invariance of aggregate attack loads preserves the defender’s incentives, but need not preserve the incentives of individual attacker types.

Our main positive result explains how exceptional this failure is in the strictly monotone simple-security-game model. Noninterchangeability forces an explicit linear equation in the prior probabilities. Equal probabilities are unnecessary, but an exact group-mass balance is necessary. We give a short proof of the aggregate-invariance lemma of [2] and recover the original multiple-resource result explicitly. These are structural results, not a polynomial-time algorithm for the computational problem in [3].

2 Model

For a nonnegative capacity vector c and a feasible budget R , write

$$\mathcal{B}(R, c) = \left\{ z : 0 \leq z_t \leq c_t \text{ for every } t, \quad \sum_t z_t = R \right\}.$$

A catcher chooses $x \in \mathcal{B}(R_0, c_0)$, and each evader $i \in [n]$ chooses $y_i \in \mathcal{B}(R_i, c_i)$. There are m sites. Let $Y_t = \sum_i y_{it}$. After dropping terms independent of a player's own strategy, the payoffs are

$$U_0(x, y) = \sum_t (\alpha_t + d_t Y_t) x_t, \quad U_i(x, y_i) = \sum_t (a_{it} - b_{it} x_t) y_{it}, \quad d_t, b_{it} > 0. \quad (1)$$

This is the catcher–evader model of [2, Section 2], with renamed parameters and strategically irrelevant terms removed. All strategy sets are assumed nonempty. A profile (x, y) is an equilibrium when every player best responds. Interchangeability here means crossing x with the *entire* evader profile y' and vice versa.

Maximizing $\sum_t f_t z_t$ over $\mathcal{B}(R, c)$ is characterized by a threshold θ :

$$z_t > 0 \implies f_t \geq \theta, \quad z_t < c_t \implies f_t \leq \theta. \quad (2)$$

Indeed, no mass can be profitably moved from a used site to a nonfull site; conversely, these inequalities imply optimality by comparing any feasible allocation with z .

Bayesian simple security games. The specialization used for our genericity results has $\alpha_t = 0$, $c_{0t} = 1$, and $R_0 = r \in \{1, \dots, m-1\}$. Type i has probability $p_i > 0$, where $\sum_i p_i = 1$, and $R_i = c_{it} = p_i$. Thus y_{it} is a *joint* probability, and y_{it}/p_i is type i 's conditional attack probability. Each realized attacker attacks one target. In the usual payoff notation,

$$d_t = u_1^1(t) - u_1^0(t), \quad a_{it} = u_2^0(i, t), \quad b_{it} = u_2^0(i, t) - u_2^1(i, t).$$

The defender's payoffs are independent of the type. The discarded defender term is $\sum_t u_1^0(t) Y_t$. Attacker best responses are exactly allocations supported on maximizers of $a_{it} - b_{it} x_t$.

Every feasible x is implementable by a distribution over r -element target sets: a vertex of $\mathcal{B}(r, 1)$ cannot have two fractional coordinates, since they admit opposite perturbations, and cannot have exactly one, since its coordinate sum is integral. Hence its vertices are precisely the incidence vectors of such sets. Equilibrium statements therefore apply to the original mixed-strategy game, not merely a relaxation. The cases $r = 0, m$ have a fixed defender strategy and are immediate. The handout [3] allows weak coverage effects; our positive theorems assume the strict inequalities in (1), while our counterexamples satisfy them.

3 Aggregate invariance and what it implies

Lemma 1 (Aggregate invariance; Li–Conitzer–Korzhyk [2], Lemma 1). *For two catcher–evader equilibria (x, y) and (x', y') ,*

$$(x'_t - x_t)(Y'_t - Y_t) = 0 \quad \text{for every } t. \quad (3)$$

If $x \neq x'$, there is a common catcher threshold λ such that

$$\alpha_t + d_t Y_t = \alpha_t + d_t Y'_t = \lambda \quad \text{whenever } x_t \neq x'_t. \quad (4)$$

Proof. Assume $x \neq x'$ and put $S_+ = \{t : x'_t > x_t\}$ and $S_- = \{t : x'_t < x_t\}$. Both are nonempty because the catcher budgets agree.

First, for each evader,

$$y'_i(S_+) \leq y_i(S_+), \quad y'_i(S_-) \geq y_i(S_-), \quad (5)$$

where $z(S) = \sum_{t \in S} z_t$. To prove the first inequality, suppose it fails. There are $t \in S_+$ and $s \notin S_+$ with $y'_{it} > y_{it}$ and $y'_{is} < y_{is}$. Write $f_{it} = a_{it} - b_{it}x_t$, and similarly for f'_{it} . Feasible transfers and optimality at the two profiles give $f_{it} \leq f_{is}$ and $f'_{is} \leq f'_{it}$. But $f'_{it} < f_{it}$ and $f'_{is} \geq f_{is}$, a contradiction. Reversing the profiles proves the second inequality.

Let θ, θ' be catcher thresholds. If $\theta' > \theta$, every $t \in S_+$ satisfies

$$\alpha_t + d_t Y_t \leq \theta < \theta' \leq \alpha_t + d_t Y'_t,$$

so $Y'_t > Y_t$ throughout S_+ , contradicting the sum of (5). Reversing the profiles rules out $\theta' < \theta$. Thus $\theta = \theta' = \lambda$.

The same threshold inequalities now imply $Y'_t \geq Y_t$ on S_+ and $Y'_t \leq Y_t$ on S_- . Their sums have the reverse inequalities by (5), so equality holds at each changed site. The threshold bounds then also give (4). $\square$

The pairwise statement (3) is equivalent to the original lemma's global formulation: at each site, either x_t or Y_t is constant across all equilibria. If two equilibria have different x_t , their common Y_t must also equal that of any third equilibrium, whose x_t differs from at least one of the first two.

Proposition 2 (One-sided interchangeability). *In any catcher–evader game, the catcher strategy from either equilibrium is a best response to the evader profile from the other.*

Proof. Set $H(Y) = U_0(x', y) - U_0(x, y)$. Equilibrium optimality gives $H(Y) \leq 0$ and $H(Y') \geq 0$, whereas Lemma 1 gives

$$H(Y') - H(Y) = \sum_t d_t (x'_t - x_t)(Y'_t - Y_t) = 0.$$

Hence both quantities vanish, proving both best-response claims. $\square$

Corollary 3 (Single evader; recovering Korzhyk–Conitzer–Parr [1], Theorem 8). *With one evader, catcher–evader equilibria are interchangeable. In particular, the non-Bayesian multiple-attacker-resource interchangeability theorem is valid under its stated assumptions.*

Proof. There is only one evader, so $Y = y_1$. Let $G(x) = U_1(x, y'_1) - U_1(x, y_1)$. Then $G(x) \leq 0$, $G(x') \geq 0$, and

$$G(x') - G(x) = - \sum_t b_{1t}(x'_t - x_t)(y'_{1t} - y_{1t}) = 0$$

by Lemma 1. Thus both crossed evader strategies best respond; Proposition 2 handles the catcher. The model of [1] has one evader, integer budgets equal to the two resource counts, and unit site capacities for both players. $\square$

4 Exact balance and generic priors

We now specialize to Bayesian simple security games. For type sets $I \subseteq [n]$ and target sets $S \subseteq [m]$, define

$$p(I) = \sum_{i \in I} p_i, \quad h(S) = \sum_{t \in S} \frac{1}{d_t}.$$

Theorem 4 (Exact balance is necessary). *If two equilibria are not interchangeable, there are disjoint nonempty type sets I, J and disjoint nonempty target sets S, T such that*

$$\frac{p(I)}{h(S)} = \frac{p(J)}{h(T)}. \quad (6)$$

More precisely, for any two equilibria with $x \neq x'$, define S_+, S_- as in Lemma 1, and let

$$v_i = \max_t (a_{it} - b_{it}x_t), \quad v'_i = \max_t (a_{it} - b_{it}x'_t), \quad I_- = \{i : v'_i < v_i\}, \quad I_+ = \{i : v'_i > v_i\}.$$

For their common threshold $\lambda \geq 0$,

$$p(I_-) = \lambda h(S_+), \quad p(I_+) = \lambda h(S_-). \quad (7)$$

If $\lambda = 0$, the two equilibria are interchangeable.

Proof. Lemma 1 gives $Y'_t = Y_t = \lambda/d_t$ on $S_+ \cup S_-$; in this specialization $\lambda \geq 0$.

If type i attacks a target in S_+ in the primed equilibrium, its utility there has strictly decreased. Since that target is now optimal, $v'_i < v_i$. Conversely, if $v'_i < v_i$, every originally optimal target must belong to S_+ : an originally optimal target whose coverage did not increase would still yield at least v_i . Therefore

$$Y'(S_+) \leq p(I_-) \leq Y(S_+). \quad (8)$$

Aggregate invariance makes the outer quantities equal. This proves the first identity in (7); reversing the profiles proves the second.

Equality throughout (8) gives a useful stronger conclusion: in *both* profiles, precisely the types in I_- place mass on S_+ , and each such type places *all* its mass there. Similarly, precisely the types in I_+ use S_- , each exclusively. All remaining types use only unchanged-coverage targets.

If $\lambda = 0$, positivity of the priors implies $I_- = I_+ = \emptyset$. Thus every type has the same optimal value at x and x' , and both attack profiles use only unchanged-coverage targets. Every used attack remains optimal after crossing. Proposition 2 supplies the defender's best-response condition.

Finally, noninterchangeability requires $x \neq x'$ and $\lambda > 0$. Equation (7) then makes I_-, I_+ nonempty, and proves (6) with $(I, J, S, T) = (I_-, I_+, S_+, S_-)$. $\square$

Corollary 5 (Generic interchangeability and uniqueness). *Fix positive defender gains $(d_t)_{t \in [m]}$. In the interior of the prior simplex, let*

$$\mathcal{H}_d = \bigcup_{\substack{I, J \subseteq [n] \text{ nonempty, disjoint} \\ S, T \subseteq [m] \text{ nonempty, disjoint}}} \{p : h(T)p(I) - h(S)p(J) = 0\}. \quad (9)$$

Outside $\mathcal{H}_d$, all equilibria are interchangeable, for every attacker payoff table with $b_{it} > 0$. Any defender-coverage differences between equilibria occur only at targets unattacked in both. If $r = 1$, the defender's equilibrium strategy is unique. The exceptional set is a finite union of proper hyperplanes, hence has measure zero and empty interior relative to the prior simplex.

Proof. Each displayed equation has a positive coefficient on I and a negative coefficient on J , so it cannot hold identically on the simplex. Theorem 4 proves interchangeability outside their finite union. For different coverage vectors, (7) forces $\lambda = 0$ outside that union, so all changed targets have zero attack load.

When $r = 1$, the catcher threshold may be chosen as $\max_t d_t Y_t > 0$, since $\sum_t Y_t = 1$. Thus two distinct equilibrium coverage vectors would force (6). This proves uniqueness. For one resource, the coverage vector is the defender's entire mixed strategy. $\square$

Membership in $\mathcal{H}_d$ is necessary, not sufficient, for failure. The uniqueness conclusion need not extend to $r > 1$. With three targets, two defender resources, one type, $d_t = 1$, and attacker utilities $(2 - x_1, -x_2, -x_3)$, every vector $(1, z, 1 - z)$, $0 \leq z \leq 1$, is an equilibrium coverage vector. The only attacked target is 1.

Corollary 6 (At least three targets and three types are necessary). *In a strictly monotone Bayesian simple security game, equilibria are interchangeable whenever $m \leq 2$ or $n \leq 2$.*

Proof. Consider equilibria with $x \neq x'$. The zero-threshold case is covered by Theorem 4. Otherwise, its proof partitions the changed targets into S_+, S_- , exclusively served by the two nonempty type groups I_-, I_+ .

If $m \leq 2$, each target group is a singleton, and its types allocate their full mass to that same target in both profiles. If $n \leq 2$, each type group is a singleton, so its type's allocation at every changed target equals the invariant aggregate load there. In either case,

$$(x'_t - x_t)(y'_{it} - y_{it}) = 0 \quad \text{for every } i, t.$$

The payoff-difference argument of Corollary 3, applied separately to each type, preserves all crossed attacker best responses. Proposition 2 finishes the proof. $\square$

5 Small counterexamples

Both examples use three targets, three types, one defender resource, and defender payoffs 1 for catching the attacker and 0 otherwise. The attacker utility matrix, with types as rows, is

$$F(x) = \begin{pmatrix} 2 - x_1 & 3 - 6x_2 & -x_3 \\ 3 - 4x_1 & 2 - x_2 & -x_3 \\ -x_1 & -x_2 & 2 - x_3 \end{pmatrix}. \quad (10)$$

All coverage penalties are strictly positive. Let e_t denote the pure attack on target t .

Equal priors: strict best responses and nonconvexity. Take $p_i = 1/3$ and

$$P = (3/8, 3/8, 1/4), \quad Q = (1/8, 1/8, 3/4), \quad A = (e_1, e_2, e_3), \quad B = (e_2, e_1, e_3).$$

Indeed,

$$F(P) = \frac{1}{8} \begin{pmatrix} 13 & 6 & -2 \\ 12 & 13 & -2 \\ -3 & -3 & 14 \end{pmatrix}, \quad F(Q) = \frac{1}{8} \begin{pmatrix} 15 & 18 & -6 \\ 20 & 15 & -6 \\ -1 & -1 & 10 \end{pmatrix}.$$

Thus A and B are the unique typewise best responses to P and Q , respectively. Both induce the uniform aggregate attack distribution, against which every defender strategy is optimal. Hence (P, A) and (Q, B) are equilibria, but neither crossing is: type 1 alone already strictly prefers to deviate in either crossing.

Moreover, $M = (P + Q)/2 = (1/4, 1/4, 1/2)$ has the unique attack best response (e_1, e_1, e_3) . Its aggregate load is $(2/3, 0, 1/3)$; defending according to M yields $1/3$, while defending target 1 surely yields $2/3$. Thus M belongs to no equilibrium, proving nonconvexity of the defender equilibrium strategy set.

Pairwise distinct priors. Keep (10), but take

$$p = (5/12, 1/4, 1/3), \quad P^* = (1/2, 1/4, 1/4), \quad Q^* = (1/5, 1/5, 3/5).$$

Use the following *conditional* attack distributions:

Type	A^* , paired with P^*	B^* , paired with Q^*
1	$(4/5, 1/5, 0)$	$(1/5, 4/5, 0)$
2	e_2	e_1
3	e_3	e_3

Type 1 is indifferent between targets 1 and 2 at both coverage vectors, with optimal utilities $3/2$ and $9/5$. Type 2 strictly prefers target 2 at P^* and target 1 at Q^* ; type 3 strictly prefers target 3 throughout. Both attack profiles again yield aggregate load $(1/3, 1/3, 1/3)$. Thus both paired profiles are equilibria. In the crossings, type 2 gains $3/4$ by deviating at P^* and $2/5$ at Q^* , respectively.

This construction extends to every prior

$$p = (q, 2/3 - q, 1/3), \quad 1/3 < q < 2/3,$$

by replacing type 1's first two attack probabilities under A^* with $(z, 1 - z)$, where $z = 1/(3q)$, and reversing them under B^* . All three priors are distinct. Nevertheless, both examples satisfy the exact balance

$$\frac{p_1 + p_2}{2} = p_3 = \frac{1}{3}.$$

They therefore lie on an exceptional hyperplane, as Theorem 4 requires. Corollary 6 shows that their numbers of targets and types are both minimal under the strict assumptions.

Verification. The accompanying standard-library Python script checks the displayed profiles and all pure deviations using exact rational arithmetic, as well as 99 instances of the unequal-prior family. It also checks the exchange step (5) on 56,258 best-response pairs in an explicitly specified finite grid, including every tie. These checks are supplementary; none of the proofs relies on numerical evidence.

A The limitation of the earlier proof

Theorem 3 of [2] is stated on page 332 of the proceedings version. Its proof appears in Section 5, page 5, of arXiv:1602.01896v2. It replaces Lemma 4 of [1] with aggregate invariance and concludes: “The remaining reasoning is unchanged.”

The distinction is between *one attacker’s allocation* and *the sum of several types’ allocations*. The earlier invariant is

$$(x'_t - x_t)(y'_t - y_t) = 0,$$

where y is the single attacker’s strategy, even when that attacker has several resources. In the Bayesian extension, only

$$(x'_t - x_t) \sum_i (y'_{it} - y_{it}) = 0$$

is available. For type i , setting $G_i(x) = U_i(x, y'_i) - U_i(x, y_i)$ gives

$$G_i(x') - G_i(x) = - \sum_t b_{it}(x'_t - x_t)(y'_{it} - y_{it}). \quad (11)$$

Aggregate invariance does not make this expression vanish. In the equal-prior example, for type 1,

$$G_1(P) = -7/24, \quad G_1(Q) = 1/8, \quad G_1(Q) - G_1(P) = 5/12.$$

Opposing changes by types 1 and 2 cancel in the aggregate, but not in either type’s incentives.

Thus the catcher-side argument corresponding to Lemma 6 of [1] extends, while its attacker-side Lemma 7 does not extend by the same reasoning. Lemma 1 and Corollary 3 independently verify the aggregate lemma and the original multiple-resource theorem. Proposition 2 gives a one-sided replacement for full interchangeability throughout the catcher–evader model; Corollary 5 supplies a generic-prior replacement for Bayesian simple security games. This correction concerns interchangeability, and by itself neither refutes the equilibrium-finding algorithm of [2] nor establishes a polynomial bound for it.

References

- [1] Dmytro Korzhyk, Vincent Conitzer, and Ronald Parr. Security games with multiple attacker resources. In *Proceedings of IJCAI*, pages 273–279, 2011. Author-hosted proceedings paper.
- [2] Yuqian Li, Vincent Conitzer, and Dmytro Korzhyk. Catcher-evader games. In *Proceedings of IJCAI*, pages 329–337, 2016. Proceedings version; full version, arXiv:1602.01896v2, April 16, 2016.
- [3] Vincent Conitzer. Open problem: Computing Bayes-Nash equilibrium in simple security games. Undated one-page problem statement. Author-hosted handout (accessed September 6, 2026).