Certified Neural Mechanisms: Finite-Sample Guarantees for Approximate IC, IR, and Fairness

Liz Lemma Future Detective

January 16, 2026

Abstract

Deep-learning approaches to mechanism design (e.g., RegretNet and its extensions surveyed in the source material) achieve impressive empirical performance but typically rely on soft penalties and lack auditable guarantees. This paper develops a clean, tractable framework for learning auctions with high-probability out-of-sample certificates for (i) approximate incentive compatibility via ex-post regret, (ii) approximate individual rationality, and (iii) fairness constraints (e.g., total-variation fairness or envy-based metrics), while maintaining feasibility by construction. We restrict attention to a structured, permutation-equivariant and Lipschitz-bounded mechanism class with differentiable feasibility layers, enabling non-vacuous uniform-convergence or PAC-Bayes style generalization bounds. Regret is estimated via adversarial best-response optimization (implemented numerically), and certification is performed on a held-out dataset with a stronger adversary to avoid systematic underestimation. The main theorem provides finite-sample bounds linking sample size, hypothesis-class complexity, and Lipschitz constants to out-of-sample regret/fairness/IR violations, and yields a revenue near-optimality guarantee within the constrained class. We validate the approach on canonical multi-item auctions and modern procurement settings (including volume-discount procurement), demonstrating that certification can be achieved with modest revenue loss relative to unconstrained deep mechanisms and substantially improved auditability for 2026-era compliance needs.

Table of Contents

3. 3. Model: multi-item auctions with additive values, feasibility constraints, and formal definitions of regret, IR violation, and fairness violation. Discussion of which parts admit closed-form structure vs require numerical procedures.

4. 4. Structured Mechanism Class: permutation-equivariant, Lipschitz-bounded networks; differentiable feasibility layers (e.g., softmax + capacity normalization; optional Sinkhorn). Boundedness and Lipschitz constants.

5. 5. Training Objective (Constrained ERM): revenue maximization subject to empirical regret/IR/fairness constraints using augmented Lagrangian; adversarial best-response inner loop (numerical) and practical approximations.

6. 6. Certification Protocol: held-out evaluation with tightened adversary; computing empirical upper bounds and translating them into high-probability population certificates; failure modes and diagnostics.

7. 7. Main Theory: uniform convergence/PAC-Bayes bounds for regret and fairness functionals; stability of the inner maximization; main theorem giving out-of-sample constraint certificates and a revenue near-optimality guarantee within the constrained class.

8. 8. Extensions: weak dependence/mixing over time; private budgets (types include B_i); dynamic settings (stagewise certificates); discussion of when closed-form mechanisms (e.g., Myerson-style) can be embedded to strengthen guarantees.

9. 9. Experiments: (i) standard multi-item multi-bidder benchmarks; (ii) fairness-constrained ad allocation; (iii) volume-discount procurement. Compare to RegretNet/EquivarianceNet/RegretFormer/ProportionNet; ablate Lipschitz/equivariance/certification adversary.

10. 10. Discussion and Policy Implications: how to report certificates, monitoring under distribution shift, compliance workflows; limitations and open problems.

11. 11. Conclusion

1 Introduction

Digital marketplaces in 2026 increasingly rely on allocation and pricing rules that are too complex to be hand-designed and too consequential to be left unaudited. Sponsored-search and retail advertising auctions now blend classical bid signals with rich contextual features (user intent, eligibility constraints, pacing, brand safety), while cloud and edge markets allocate heterogeneous compute under latency and carbon constraints. Across these domains, platforms face a familiar economic tension: we want mechanisms that are expressive enough to monetize complex demand and enforce business rules, yet predictable enough that strategic participants can trust the rules and regulators can verify compliance. This tension has pushed the field toward *learned mechanisms*—parameterized allocation and payment maps trained from data—because learning offers a principled way to incorporate high-dimensional context, soft constraints, and objectives that are hard to encode analytically.

At the same time, learned mechanisms raise a distinct governance problem: *auditable correctness*. In classical mechanism design, incentive and individual-rationality guarantees are derived symbolically from an explicit form (e.g., VCG, Myerson), and feasibility is ensured by construction. In learned systems, by contrast, the designer often selects a hypothesis class (typically neural), optimizes an empirical objective, and then deploys the resulting model as a black box. If we only report training curves (revenue up, regret down), we have not produced a statement that is meaningful for an auditor, a court, or even internal risk management. What is needed is an explicit bridge from *empirical* performance on sampled valuation profiles to *population* guarantees under the (unknown) environment generating bids and contexts. The motivating question is therefore not merely "can we train a high-revenue approximately truthful network on a benchmark?" but rather: *can we document, with quantifiable uncertainty, how far the deployed mechanism can deviate from incentive compatibility, individual rationality, and fairness constraints on future market instances?*

A large and influential line of work has made learned mechanisms practical by introducing differentiable surrogates for incentive constraints and training them with penalties or Lagrange multipliers. Systems in the spirit of RegretNet and its variants operationalize approximate dominant-strategy incentive compatibility by sampling valuation profiles, computing (approximate) best-response misreports, and penalizing the resulting empirical regret in the training objective. Subsequent designs incorporate additional side constraints, including budget feasibility, proportionality or exposure constraints, and group-level fairness regularizers; more recent architectures add symmetry (permutation equivariance), attention, or transformer-based components to scale to larger bidder sets. From an engineering perspective, the appeal is clear: one can train a flexible mechanism end-to-end using stochastic gradi-

ents, and the output often looks well-behaved on the training distribution.

However, penalty-only training leaves a gap between what we *optimize* and what we can *certify*. The first issue is statistical: empirical regret and empirical fairness are estimates computed from a finite sample, and without further control they can be severely optimistic. A mechanism class rich enough to fit complicated patterns in context can also fit idiosyncrasies of the training sample. When regret is computed via an inner maximization over misreports, this optimism can be amplified: the training loop adapts the mechanism parameters to the same sample used to evaluate constraints, while the misreport search is itself a noisy numerical procedure. As a result, a small reported regret at the end of training may reflect (i) overfitting to the sampled profiles, (ii) under-optimized best responses, or (iii) fragile cancellations that disappear out of sample. The second issue is operational: compliance teams typically require a *worst-case* or high-confidence bound, not a point estimate. Hyperparameter choices (penalty weights, learning rates, oracle iterations) are rarely motivated by explicit risk tolerances, and two runs with similar empirical metrics can have very different tail behavior. The third issue is distributional: real markets drift. Even if the data are approximately i.i.d. in a lab setting, deployment introduces time dependence (seasonality, bidder learning, platform updates), and the relevant object is performance under the ongoing data-generating process, not on a static dataset.

These limitations matter precisely because the constraints we care about are normative, not merely predictive. Approximate incentive compatibility limits the platform's ability to extract value through opaque manipulation and stabilizes participation by reducing strategic uncertainty. Approximate individual rationality ensures that participation does not systematically harm bidders (or downstream stakeholders) ex post, which is increasingly salient when small firms or regulated advertisers participate. Fairness constraints—whether interpreted as group exposure, user-level similarity, or envy-based surrogates—are often motivated by legal or contractual commitments and thus require documentation. In short, we do not merely want mechanisms that *tend* to be truthful and fair on sampled instances; we want mechanisms whose violations are *bounded* in a way that is legible to auditors.

Our view is that this calls for importing a familiar idea from learning theory and safety-critical ML into mechanism design: *certificates*. A certificate is a computable bound that converts held-out empirical evaluations into high-confidence statements about population performance, with explicit dependence on sample size, model complexity, and numerical approximation error. In the present setting, the certificate should upper-bound incentive violations (via regret), individual-rationality violations, and fairness violations simultaneously, and it should do so for the *learned* mechanism selected by the training procedure. Practically, this allows a platform to say: "With probability at least $1-\delta$ over the sampled market instances used for certifica-

tion, the deployed mechanism's expected regret is at most ε_r (and similarly for IR and fairness)," where ε_r decomposes into an empirical estimate plus a slack term that shrinks with more data and tighter architectural control.

The key conceptual move is to treat the relevant quantities—regret integrands, IR shortfalls, fairness penalties—as *functionals* induced by the mechanism class, and then to bound their generalization error uniformly over that class. This immediately clarifies why purely penalty-based reports are insufficient: without controlling the effective complexity of the mechanism class and the stability of the inner maximization, there is no reason for empirical constraint satisfaction to persist out of sample. It also clarifies why certain architectural choices, often justified heuristically, are central to auditability. Feasibility-by-construction removes an entire category of violations from the certification problem. Permutation equivariance (treating bidders symmetrically absent distinguishing context) both matches economic symmetry and reduces statistical complexity. Lipschitz control in valuations limits how sharply utilities can change with small bid perturbations, which stabilizes both regret estimation and uniform convergence bounds. These are not merely technical conveniences; they are design principles that convert a black-box learner into a mechanism that is, in a precise sense, *auditable*.

We also emphasize a limitation that must be made explicit to avoid overclaiming: certificates do not eliminate the need for a best-response oracle, and they cannot certify properties that are not encoded in the evaluation functional. If fairness is defined at the level of allocation exposure, we can certify that notion, but not necessarily other legal notions of discrimination. If the oracle only finds approximately optimal misreports, then any regret estimate inherits an "oracle gap" that must be accounted for. Our goal is therefore not to present a frictionless path to perfect truthfulness or perfect fairness, but to provide a disciplined accounting framework: what must be assumed (boundedness, Lipschitzness, symmetry), what is measured empirically (held-out regret/IR/fairness), what is approximated numerically (best responses), and how these ingredients combine into an end-to-end bound that can be communicated to non-technical stakeholders.

Finally, positioning certificates within the economic logic highlights the central trade-off. Constraining regret, IR violations, and fairness violations restricts the feasible set of mechanisms and can reduce revenue relative to an unconstrained optimum. Penalty-based training already navigates this trade-off implicitly through hyperparameters; our contribution is to make the trade-off *explicit and accountable* by tying it to user-chosen tolerances $(\varepsilon_r, \varepsilon_{ir}, \varepsilon_f)$ and to finite-sample uncertainty. This perspective also suggests a practical workflow: train with constraints on a training set, then certify on an independent holdout (and, where appropriate, under stress tests reflecting plausible deviations), and re-certify periodically as the market evolves. In that sense, the model we develop is not only a learning algorithm but also a governance protocol: it illuminates how a platform can deploy learned

mechanisms while producing the documentation that modern markets increasingly require.

2 Related Work

Our paper sits at the intersection of three literatures: (i) deep learning approaches to mechanism design, (ii) learning-theoretic analyses of constrained and bilevel objectives (especially those involving best-response computations), and (iii) algorithmic auditing and compliance frameworks that motivate high-confidence, stakeholder-facing guarantees.

Deep learning for mechanism design. A first thread learns auctions by parameterizing allocation and payment rules with neural networks and training them to optimize expected revenue (or welfare) subject to approximate incentive and participation constraints. The most widely used approach in multi-dimensional settings follows the "regret minimization" template popularized by RegretNet and successors **???**. The key idea is pragmatic: dominant-strategy incentive compatibility (DSIC) is replaced with an ex-post regret objective estimated on samples, where regret itself is computed by (approximately) maximizing a bidder's utility over misreports. Training then proceeds by stochastic gradient methods, often using Lagrangian or penalty formulations for regret and individual rationality (IR). From an economic perspective, these methods trade exactness for flexibility: rather than restricting attention to analytically tractable families (e.g., affine maximizers), they allow the mechanism to respond to rich contextual features and high-dimensional types, at the cost of relying on numerical oracles and finite-sample evaluations.

A complementary line enforces IC using *structural* characterizations that avoid explicit best-response search. Rochet's classical result characterizes implementable allocation rules via cyclic monotonicity and convex potentials; several neural architectures exploit this connection by representing utilities or allocations through convex networks, thereby guaranteeing IC (or approximate IC) by construction **??**. These methods are attractive when one can parameterize the relevant convex object at scale, but they can be restrictive in environments with many items, contextual constraints, or fairness objectives that couple allocations across agents or users. In practice, many marketplace deployments therefore blend structural ideas (e.g., feasibility layers) with regret-based training to preserve expressiveness.

Architectural work has also focused on the symmetries inherent in anonymous environments. In standard auctions, bidders are exchangeable ex ante, and a mechanism that hard-codes permutation equivariance is both economically natural and statistically advantageous. Several designs incorporate equivariant layers (DeepSets, attention with shared parameters, or

graph neural networks) to ensure that relabeling bidders merely relabels outputs **???**. Empirically, these architectures often improve sample efficiency and out-of-distribution behavior, and conceptually they reduce the effective hypothesis-class complexity in precisely the way that matters for generalization bounds. Recent work further explores transformer-style mechanisms that scale to larger bidder sets and capture cross-bidder interactions through attention, sometimes under names such as "RegretFormer" or related attention-based learned auctions **??**. While these models can be highly expressive, they also underscore the need for principled control of sensitivity and complexity if one wants auditable constraint satisfaction rather than merely low training regret.

Fairness constraints and redistribution. A second set of contributions extends learned mechanism design to incorporate normative or policy-driven constraints beyond IC/IR. In ad auctions and recommendation-like allocations, fairness is often operationalized as exposure parity across advertiser categories, user-level similarity constraints, or bounded disparity in allocation probabilities across protected groups **??**. In the learned-mechanism setting, these notions are typically imposed via differentiable penalties on allocation vectors (or on downstream exposure outcomes), sometimes coupled with constraints on payments, budgets, or pacing **??**. Economically, such constraints reflect that platforms are multi-objective: revenue is traded off against contractual commitments, user experience, and legal risk. Methodologically, fairness constraints are challenging because they are often *global* (depending on distributions over contexts) and can be sensitive to sampling noise, which again points toward the need for held-out evaluation and explicit uncertainty quantification.

Related but distinct is the literature on *redistribution* and budget-balance constraints. In classical mechanism design, payments are central for incentives but also raise concerns about surplus extraction, collusion, and regulatory acceptability. Learned approaches have incorporated payment caps, rebates, or redistribution layers to control how revenue is collected and potentially returned, drawing inspiration from redistribution mechanisms and budget-feasible design **??**. These extensions strengthen the case that the relevant constraints are not limited to IC/IR: real systems are constrained optimization problems whose feasible set is shaped by business and policy requirements.

Learning theory for constrained and bilevel optimization. The training objectives used in regret-based learned mechanisms are inherently bilevel: an outer loop optimizes mechanism parameters, while an inner loop computes (approximate) best responses. This places the problem close to modern bilevel learning in adversarial training and robust optimization, where

the quantity of interest is often a supremum over perturbations or adversarial strategies **?**. In mechanism design, the inner maximization has a clear economic meaning (a bidder deviation), and the gap between the computed best response and the true best response has a direct interpretation as under-estimated strategic vulnerability. Several papers study gradient estimators, differentiable approximations, and convergence heuristics for this inner problem **??**. From a certification standpoint, however, the key issue is not only optimization performance but also how oracle error propagates into guarantees—a point that motivates our explicit accounting for best-response approximation.

On the statistical side, there is a large learning-theory literature on constrained empirical risk minimization, uniform convergence, and generalization under complexity control **??**. Our focus aligns with work that treats constraint violations as expectations of bounded functionals and then derives high-probability bounds via Rademacher complexity, covering numbers, or PAC-Bayes techniques **??**. In our setting, the relevant function classes are induced by *mechanism-induced* quantities such as regret integrands and fairness penalties; bounding their complexity requires using architectural properties (equivariance, Lipschitz control, bounded outputs) rather than only parameter counts. There is also a related literature on generalization under dependence (mixing processes) that is relevant for repeated-market data, where observations can be temporally correlated **??**. While many learned-auction experiments assume i.i.d. samples, deployment in ad markets or cloud markets naturally produces time dependence, and any auditing-oriented analysis should at least acknowledge how certificates degrade (or can be repaired) under weak dependence.

Algorithmic auditing and compliance. Finally, our motivation is closely connected to work on algorithmic accountability, which emphasizes that empirical performance alone is not a compliance artifact. In domains such as lending, hiring, and advertising, regulators and internal governance teams increasingly require documented evidence of constraint satisfaction, monitoring protocols, and uncertainty quantification **??**. Technical work on algorithmic auditing develops procedures for black-box testing, stress testing under distribution shift, and post-deployment monitoring, including approaches that resemble "certificates" in spirit: high-confidence bounds, conservative estimates, and explicit failure probabilities **??**. In machine learning more broadly, conformal prediction and related methods provide distribution-free coverage guarantees for predictive uncertainty **?**; while the object differs (prediction sets rather than incentive violations), the governance logic is similar: stakeholders need interpretable, probabilistic guarantees that survive deployment uncertainty.

We view certified learned mechanisms as importing this auditing perspec-

tive into economic design. Compared to prior learned-mechanism work that reports regret and fairness on the training distribution, our emphasis is on producing a statement an auditor can read: a bound that decomposes into (i) what was measured on held-out data, (ii) what is owed to finite-sample uncertainty and model complexity, and (iii) what is owed to numerical approximation in best-response computation. This orientation does not replace the rich algorithmic contributions of the learning-based mechanism design literature; rather, it reframes them through the lens of accountability, highlighting which architectural and statistical choices make a learned mechanism not only high-performing, but also governable.

3 Model

We study a contextual, multi-item auction environment with strategic bidders and an auctioneer (or platform) who chooses a mechanism from a parameterized class. There are $n \geq 1$ bidders (agents) indexed by $i \in \{1, \ldots, n\}$ and $m \geq 1$ items indexed by $\ell \in \{1, \ldots, m\}$. Each bidder i has an additive valuation vector $v_i \in [0,1]^m$, where $v_{i\ell}$ is bidder i's value for item ℓ. We write the valuation profile as $v = (v_1, \ldots, v_n) \in \mathcal{V} \subseteq [0,1]^{n \times m}$. We also allow the mechanism to condition on an observed context $x \in \mathcal{X}$, which can include user or query features, eligibility constraints, reserve policies, business rules, or other market covariates. The pair (v, x) is drawn from an unknown distribution $\mathcal{D}$; in an offline-learning interpretation we observe a sample $\{(v^{(s)}, x^{(s)})\}_{s=1}^N$ drawn i.i.d. from $\mathcal{D}$, while in repeated-market settings we allow for weak temporal dependence (e.g., β-mixing) and interpret $\mathcal{D}$ as the stationary distribution.

A (possibly randomized) direct-revelation mechanism consists of an allocation rule and a payment rule. Bidders submit bid reports $b_i \in [0,1]^m$ and $b = (b_1, \ldots, b_n)$. The mechanism outputs (i) an allocation-probability matrix $p(b, x) \in [0,1]^{n \times m}$, where $p_{i\ell}(b, x)$ is the probability (or fractional share) with which bidder i receives item ℓ, and (ii) a payment vector $t(b, x) \in \mathbb{R}_+^n$, where $t_i(b, x)$ is bidder i's payment. We use the probabilistic interpretation for two reasons. First, many real systems are randomized (e.g., via tie-breaking, pacing, or exploration). Second, even when the deployed mechanism is deterministic, relaxing to fractional allocations is analytically convenient and aligns with standard neural parameterizations that output soft assignment weights. Throughout, feasibility constraints ensure that the allocation respects per-item capacity:

$$0 \leq p_{i\ell}(b, x) \leq 1, \qquad \sum_{i=1}^{n} p_{i\ell}(b, x) \leq 1 \quad \forall \ell \in \{1, \ldots, m\}. \tag{1}$$

The slack $\left(1 - \sum_i p_{i\ell}\right)$ can be interpreted as the probability that item ℓ remains unallocated (equivalently, allocation to a dummy bidder). We em-

phasize that we do *not* impose a unit-demand constraint on bidders; with additive values a bidder can simultaneously receive multiple items, subject only to per-item capacity.[1]

Given true values v_i and context x, bidder i's quasi-linear utility under report profile b is

$$u_i(b; x, v_i) \;=\; \sum_{\ell=1}^{m} p_{i\ell}(b, x)\, v_{i\ell} \;-\; t_i(b, x). \tag{2}$$

We take additivity and quasilinearity as the baseline model for many multi-slot and multi-product allocations (notably in ads and sponsored content), and we normalize values to $[0, 1]$ to make sensitivity and concentration statements scale-free. The platform's revenue under a mechanism (p, t) is the expected sum of payments,

$$\mathrm{REV} \;=\; \mathbb{E}_{(v,x)\sim\mathcal{D}}\Big[\sum_{i=1}^{n} t_i(v, x)\Big], \tag{3}$$

where, for direct mechanisms, we identify truthful reports with bids (i.e., $b = v$) when defining the objective. In deployment, bidders may deviate, and the relevant question is how costly such deviations can be for incentives and participation.

Incentives via ex-post regret. Our incentive benchmark is dominant-strategy incentive compatibility (DSIC): truthful reporting should maximize utility regardless of others' bids and context. In multi-dimensional environments with contextual constraints, exact DSIC is typically difficult to guarantee without restrictive structure. We therefore work with *ex-post regret* as an economically meaningful proxy for IC violations. For a fixed mechanism (p, t), the per-instance gain from deviation is

$$\phi_i(v, x) \;=\; \max_{v_i' \in [0,1]^m} u_i\big((v_i', v_{-i}); x, v_i\big) \;-\; u_i\big((v_i, v_{-i}); x, v_i\big),$$

and the population regret is the expectation

$$\mathrm{RGT}_i \;=\; \mathbb{E}_{(v,x)\sim\mathcal{D}}\big[\phi_i(v, x)\big]. \tag{4}$$

When $\mathrm{RGT}_i = 0$ for all i, truthful reporting is a best response almost surely, yielding DSIC. Positive regret quantifies the maximum utility improvement from misreporting, and thus has a direct interpretation for auditing: it upper-bounds the incentive to manipulate the mechanism, measured in the same units as utility (normalized value).

[1] Additional feasibility constraints (e.g., budgets, frequency caps, matroid constraints) can be incorporated, but we focus on (1) to isolate the strategic and statistical issues.

A crucial modeling point is that the inner maximization over $v_i' \in [0,1]^m$ is generally non-convex once $p(\cdot, x)$ and $t(\cdot, x)$ are represented by expressive function approximators. In contrast to classical single-parameter auctions (where monotonicity and payment identities yield closed-form computations of incentives), here we should *expect* to rely on numerical best-response search. Operationally, given a sample (v, x) and mechanism parameters, we approximate the best response by running a gradient-based or derivative-free optimizer over the bounded domain $[0,1]^m$ (e.g., projected gradient ascent with multi-start). This is not merely a computational detail: the quality of regret estimates depends on how well we solve this inner problem, and any under-optimization directly translates into underestimated strategic vulnerability. For this reason, we treat best-response computation as an explicit oracle whose approximation error can be tracked and, when needed, folded conservatively into the final guarantees.

Participation via IR violations. Beyond incentives, platforms and regulators often require participation guarantees. In a quasilinear setting, ex-post individual rationality (IR) requires that truthful utility be nonnegative for every realization:

$$u_i\big((v_i, v_{-i}); x, v_i\big) \;\geq\; 0 \qquad \text{for all } (v, x).$$

Because learned mechanisms may violate IR on rare contexts or for certain types (especially when optimizing revenue), we quantify participation risk through an *IR violation functional*

$$\mathrm{IRV}_i \;=\; \mathbb{E}_{(v,x)\sim\mathcal{D}}\Big[\max\big\{0,\, -u_i\big((v_i, v_{-i}); x, v_i\big)\big\}\Big]. \tag{5}$$

This measures the expected magnitude of negative utility (rather than merely its probability), which aligns with risk management: occasional small violations may be acceptable (or remediable via refunds), while large violations are typically unacceptable. As with regret, IRV is an expectation under $\mathcal{D}$ and is therefore naturally estimated from samples, with the caveat that tail events may require additional stress testing if the deployment distribution is believed to shift.

Fairness as a population constraint. Many marketplace deployments include normative constraints that cannot be reduced to individual incentives. We model such requirements via a fairness violation functional FR that depends (possibly nonlinearly) on allocations and contexts and is evaluated in expectation under $\mathcal{D}$. Importantly, fairness constraints are often *global*: they compare allocations across users, across bidder groups, or across contexts, and thus cannot be validated from a single instance alone without reference to a broader distributional baseline.

To keep the framework flexible, we allow FR to be any bounded functional of the allocation rule (and context) that is computable on data. A concrete example, motivated by total-variation style individual fairness, is the following. Let $\mathcal{U}$ be a finite set of user contexts (or a discretization thereof), let advertisers be partitioned into classes $\{C_c\}$ (e.g., protected groups or categories), and let $d_c(u, u')$ be an allowed disparity between users u and u'. Define a per-user unfairness penalty

$$\mathrm{Unf}_u \;=\; \sum_{u' \in \mathcal{U}} \sum_{c} \max\Big\{0,\; \sum_{i \in C_c} \big|p_i(u) - p_i(u')\big| - d_c(u, u')\Big\},$$

where $p_i(u)$ denotes bidder i's allocation probability (possibly aggregated over items) under user context u. Then the fairness violation is

$$\mathrm{FR} \;=\; \mathbb{E}\big[\mathrm{Unf}_u\big], \tag{6}$$

with the expectation taken over the user distribution induced by $\mathcal{D}$. Variants include exposure parity constraints, envy or dissatisfaction penalties, and EF1-style surrogates. The economic point is that such constraints formalize policy and product requirements in the same language as incentives: as expectations of measurable functionals that can be estimated on held-out data and audited with explicit uncertainty.

What is structural and what is numerical. Our modeling choices deliberately separate three layers of difficulty. First, feasibility (1) is a hard constraint that we will enforce deterministically (so it does not rely on sample averages). Second, IC, IR, and fairness are expressed as expectations (4)–(6), which are statistically estimable but require generalization control: low empirical violation does not automatically imply low population violation. Third, computing the regret integrand requires solving an inner optimization over misreports; unlike in analytically tractable auction families, this step is inherently numerical in expressive contextual settings. From a practice and governance perspective, this decomposition clarifies what an auditor should ask for: (i) architectural or procedural evidence that feasibility cannot be violated at runtime, (ii) high-confidence bounds translating empirical IC/IR/fairness measurements into population guarantees, and (iii) documentation of the best-response search procedure (including its approximation error) to ensure that incentives were not assessed with an overly weak adversary.

4 Structured mechanism class

To make the learning-and-certification problem well posed, we restrict attention to a hypothesis class Θ that builds in the two properties an auditor

can (and should) demand as *structural*: feasibility and symmetry. The remaining requirements—small regret, small IR violations, and small fairness violations—are then treated as *statistical* properties of the learned mechanism that must generalize out of sample. At a high level, our design principle is that architectural structure should remove "runtime failure modes" (over-allocation, label dependence), while regularity (Lipschitz control and bounded outputs) should make the empirical estimates stable enough to support high-confidence certificates.

A parameterized direct mechanism. We consider mechanisms parameterized by $\theta \in \mathbb{R}^d$ that map bid reports and context into allocations and payments:

$$(p_\theta, t_\theta) : \ [0,1]^{n\times m} \times \mathcal{X} \to [0,1]^{n\times m} \times \mathbb{R}^n_+.$$

We keep the discussion in terms of bids b (rather than values) because both training and regret computation require evaluating the mechanism under arbitrary misreports. In direct-revelation training we will evaluate objectives at $b = v$, but the mechanism itself must be defined for all $b \in [0,1]^{n\times m}$.

Permutation equivariance as a modeling commitment. Because bidder identities are economically meaningless labels, we impose permutation equivariance in bidders: for any permutation $\pi \in \Pi_n$ and any input (b, x),

$$p_\theta(\pi b, x) = \pi\, p_\theta(b, x), \qquad t_\theta(\pi b, x) = \pi\, t_\theta(b, x), \tag{7}$$

where $(\pi b)_i = b_{\pi^{-1}(i)}$ and $(\pi p)_{i\ell} = p_{\pi^{-1}(i)\ell}$. This restriction has a practical and a statistical interpretation. Practically, it prevents arbitrary dependence on the indexing of bidders (a clear governance failure mode in deployments where bidders enter and exit). Statistically, it reduces effective hypothesis-class complexity by tying together the behavior across relabelings, improving sample efficiency in exactly the way exchangeability suggests. When items are homogeneous (or partitioned into exchangeable classes), the same idea can be applied to item permutations; we leave item equivariance optional because many applications have item-specific semantics (e.g., distinct ad slots or heterogeneous products).

A canonical way to implement (7) is with DeepSets-style architectures. For example, letting ϕ embed each bidder's bid vector (and context) and letting $\oplus$ denote a permutation-invariant aggregator (sum/mean), we can form bidder representations

$$h_i \ = \ \rho\Big(\phi(b_i, x), \ \oplus_{j=1}^n \phi(b_j, x)\Big), \tag{8}$$

with shared parameters in ϕ and ρ across bidders. More expressive variants replace the single aggregate with multi-head self-attention with shared weights, which is also permutation equivariant when it operates on sets of

bidder embeddings. The key point is not the particular neural primitive, but that (7) is enforced by construction rather than encouraged by a penalty.

Differentiable feasibility layers for allocations. We parameterize allocations by first producing unconstrained scores (logits) and then projecting them into the feasible region using a differentiable normalization layer. Concretely, let $s_\theta(b, x) \in \mathbb{R}^{n \times m}$ be a score matrix produced by an equivariant network (e.g., using (8) and then decoding item-wise). For each item ℓ, we define allocation probabilities by a softmax across bidders,

$$p_{\theta,i\ell}(b, x) \;=\; \frac{\exp(s_{\theta,i\ell}(b, x))}{\sum_{j=0}^{n} \exp(s_{\theta,j\ell}(b, x))}, \tag{9}$$

where $j = 0$ denotes a dummy bidder capturing unallocated probability mass. Then $p_{\theta,i\ell} \in [0, 1]$ and $\sum_{i=1}^{n} p_{\theta,i\ell} \leq 1$ for all inputs, deterministically. This is the simplest instance of "feasibility-by-construction" and is attractive for two operational reasons: (i) feasibility is not something we must estimate (there is no statistical uncertainty), and (ii) the mapping remains smooth, which is essential for gradient-based training and for gradient-based best-response computation in regret estimation.

When feasibility constraints are more structured than per-item capacity, one can replace (9) with a differentiable projection onto a richer polytope. A common choice is a Sinkhorn normalization layer that approximately enforces doubly-stochastic constraints (row and column sums), which is useful when items must be fully allocated, when bidders have unit-demand style capacity constraints, or when exposure budgets are imposed. In such cases, we interpret the normalization as an approximate projection operator $\mathcal{P}$ and set $p_\theta = \mathcal{P}(s_\theta)$; the approximation error is then part of the numerical pipeline, and in a conservative certification one can add a small feasibility slack or include a dummy option to preserve hard constraints.

Payments: nonnegativity and boundedness. We similarly encode basic payment constraints architecturally. Because we allow randomized allocations and focus on ex-post utility and regret, we do not rely on closed-form payment identities; instead, we learn a payment rule t_θ jointly with p_θ and certify its incentive properties via regret. However, two restrictions are economically natural and statistically useful: payments should be nonnegative and uniformly bounded. We therefore parameterize

$$t_{\theta,i}(b, x) \;=\; B \cdot \sigma\big(g_{\theta,i}(b, x)\big), \tag{10}$$

where g_θ is an equivariant network, $\sigma(z) = (1 + e^{-z})^{-1}$ is the logistic function, and $B < \infty$ is a design bound. This ensures $t_{\theta,i} \in (0, B)$ for all inputs. The upper bound is not merely technical: it prevents the learning problem from creating extremely large transfers on rare contexts (which

would inflate regret and IRV tails) and it improves concentration when we estimate expectations from data. In applications where exact zero payments matter (e.g., to allow non-winners to pay exactly 0), one can use $t_{\theta,i} = B \cdot \mathrm{ReLU}(g_{\theta,i})/(1 + \mathrm{ReLU}(g_{\theta,i}))$, or explicitly multiply (10) by an allocation-dependent mask; the central requirement for our theory is boundedness, not the particular squashing map.

Lipschitz control and why we impose it. The certificate step requires uniform convergence of regret, IRV, and fairness functionals over $\theta \in \Theta$. In expressive neural classes, uniform convergence can fail without some form of complexity control. We adopt a regularity condition that is both interpretable and enforceable: uniform Lipschitzness of the mechanism in bids (equivalently in values on the truthful path). Formally, we assume that for all $\theta \in \Theta$, all contexts x, and all bid profiles b, b',

$$\big\|p_\theta(b,x) - p_\theta(b',x)\big\|_1 \;+\; \big\|t_\theta(b,x) - t_\theta(b',x)\big\|_1 \;\le\; L\,\|b - b'\|_1, \qquad (11)$$

for a uniform constant $L < \infty$. Economically, (11) rules out mechanisms where a bidder can induce discontinuous jumps in allocation or payment by arbitrarily small perturbations of a report. Such jumps are precisely what make regret estimation brittle: the inner maximization over misreports becomes sensitive to numerical optimization and to sampling noise. Statistically, (11) makes the per-sample regret integrand and fairness penalties Lipschitz functions of (v, x) (under mild regularity), which is the key input into Rademacher- or covering-number bounds for certificates.

In practice, (11) is implemented by combining three ingredients. First, we bound intermediate activations and logits, for instance by using tanh at the last layer producing scores $s_{\theta,i\ell} \in [-S, S]$, which also controls the Jacobian of the softmax map in (9). Second, we constrain weight matrices via spectral normalization or weight clipping, yielding explicit bounds on the Lipschitz constants of the constituent linear maps. Third, we use 1-Lipschitz nonlinearities (e.g., ReLU, leaky-ReLU, group norm without learned scale) or track their Lipschitz factors when they exceed 1. Under these design choices, one can compute an a priori upper bound on L as a product (or sum, in residual architectures) of layer-wise operator norms, and treat the resulting L as part of the mechanism documentation provided for auditing.

Putting the class together. We summarize the structured class as

$$\Theta \;=\; \Big\{\theta :\; (p_\theta, t_\theta) \text{ satisfies (7), feasibility by construction via (9) (or a differentiable projection), bou}$$

This definition makes explicit what is guaranteed at runtime (feasibility, symmetry, bounded transfers) and what must be learned and certified (low regret, low IRV, low FR). It also clarifies a limitation: the more tightly we

enforce Lipschitzness and symmetry, the smaller the hypothesis class and the more conservative the revenue frontier may become. We view this as an economically meaningful tradeoff rather than a purely statistical artifact: mechanisms that are extremely sensitive to reports are harder to justify from a governance perspective, precisely because they are harder to audit and easier to manipulate.

Finally, we note that the constant L we track is typically an upper bound and may be loose. This looseness affects the sharpness of worst-case certificates (through complexity terms), but it does not invalidate the basic logic: bounding sensitivity is what allows us to translate held-out measurements of regret, IR, and fairness into statements that remain reliable under deployment sampling variation.

5 Training objective: constrained ERM with adversarial regret

Having fixed a structured hypothesis class Θ in which feasibility and symmetry hold by construction, we now treat incentive compatibility, individual rationality, and fairness as *out-of-sample* requirements that must be learned from data and later certified. Our training problem is therefore a constrained empirical risk minimization (ERM) problem: we choose parameters θ to maximize empirical revenue while keeping empirical proxies for regret, IR violations, and fairness violations below prespecified tolerances.

Data split and empirical objectives. We observe samples $\{(v^{(s)}, x^{(s)})\}_{s=1}^{N}$ from $\mathcal{D}$ (i.i.d. or weakly dependent). Because the subsequent certificate is only meaningful when computed on data not used to tune θ, we conceptually split the data into a training set $\mathcal{S}_{\mathrm{tr}}$ and a holdout set $\mathcal{S}_{\mathrm{ho}}$ (the latter will be used in Section 6). In this section we focus on training on $\mathcal{S}_{\mathrm{tr}}$ of size N_{tr}.

On the truthful path we evaluate the mechanism at $b = v$, but the regret computation will explicitly query the mechanism at misreports. Given θ, define the empirical revenue

$$\widehat{\mathrm{REV}}_{\mathrm{tr}}(\theta) \;=\; \frac{1}{N_{\mathrm{tr}}} \sum_{(v,x)\in\mathcal{S}_{\mathrm{tr}}} \sum_{i=1}^{n} t_{\theta,i}(v,x).$$

The empirical IR-violation estimate is straightforward because it depends only on truthful utility:

$$\widehat{\mathrm{IRV}}_{\mathrm{tr},i}(\theta) \;=\; \frac{1}{N_{\mathrm{tr}}} \sum_{(v,x)\in\mathcal{S}_{\mathrm{tr}}} \max\Big\{0,\; -u_i((v_i, v_{-i}); x, v_i)\Big\}.$$

For fairness, we assume we have a per-sample (or per-mini-batch) measurable penalty $\psi_\theta(v,x)$ whose expectation is $\mathrm{FR}(\theta)$ (e.g., total-variation violations

computed by pairing contexts within a batch). We then define

$$\widehat{\mathrm{FR}}_{\mathrm{tr}}(\theta) \;=\; \frac{1}{N_{\mathrm{tr}}} \sum_{(v,x)\in\mathcal{S}_{\mathrm{tr}}} \psi_\theta(v,x).$$

Empirical regret as a bilevel objective. The central complication is regret. For each sample (v, x) and bidder i, define the sample-level best-response value

$$\phi_{i,\theta}(v,x) \;=\; \max_{b_i\in[0,1]^m} \; u_i\big((b_i, v_{-i}); x, v_i\big) \;-\; u_i\big((v_i, v_{-i}); x, v_i\big),$$

and the empirical regret estimate

$$\widehat{\mathrm{RGT}}_{\mathrm{tr},i}(\theta) \;=\; \frac{1}{N_{\mathrm{tr}}} \sum_{(v,x)\in\mathcal{S}_{\mathrm{tr}}} \phi_{i,\theta}(v,x).$$

Computationally, $\widehat{\mathrm{RGT}}_{\mathrm{tr},i}(\theta)$ is a bilevel quantity: an outer expectation over samples and an inner maximization over misreports. We deliberately work with this ex-post regret proxy (rather than an ex-interim IC constraint) because it is agnostic to the mechanism's parametric form and remains meaningful in the presence of contextual features x and randomized allocations.

Constrained ERM formulation. Fix target tolerances $(\varepsilon_r, \varepsilon_{ir}, \varepsilon_f)$ for regret, IR violation, and fairness, respectively. The training problem is

$$\begin{aligned}
\max_{\theta\in\Theta} \quad & \widehat{\mathrm{REV}}_{\mathrm{tr}}(\theta) \\
\text{s.t.} \quad & \widehat{\mathrm{RGT}}_{\mathrm{tr},i}(\theta) \le \varepsilon_r, \qquad i\in[n], \\
& \widehat{\mathrm{IRV}}_{\mathrm{tr},i}(\theta) \le \varepsilon_{ir}, \qquad i\in[n], \\
& \widehat{\mathrm{FR}}_{\mathrm{tr}}(\theta) \le \varepsilon_f.
\end{aligned} \tag{12}$$

Although (12) is written with hard empirical constraints, in practice we will not solve it exactly. Instead, we use a smooth constrained optimization scheme that (i) can be implemented with stochastic gradients, and (ii) produces mechanisms whose empirical violations are small enough that the holdout certificate (Section 6) can plausibly clear them with statistical slack.

Augmented Lagrangian training. A convenient approach is an augmented Lagrangian (or, equivalently, a primal–dual penalty method) in which we penalize constraint violations while maintaining explicit dual variables. Let $\lambda = (\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n_+$ be multipliers for regret, $\mu \in \mathbb{R}^n_+$ for IR violations, and $\nu \in \mathbb{R}_+$ for fairness. For a penalty weight $\rho > 0$ and hinge

notation $[z]_+ = \max\{0, z\}$, define

$$\mathcal{L}_\rho(\theta, \lambda, \mu, \nu) \;=\; -\widehat{\mathrm{REV}}_{\mathrm{tr}}(\theta) + \sum_{i=1}^{n} \lambda_i \big(\widehat{\mathrm{RGT}}_{\mathrm{tr},i}(\theta) - \varepsilon_r\big) + \sum_{i=1}^{n} \mu_i \big(\widehat{\mathrm{IRV}}_{\mathrm{tr},i}(\theta) - \varepsilon_{ir}\big) + \nu\big(\widehat{\mathrm{FR}}_{\mathrm{tr}}(\theta) - \varepsilon_f\big)$$
$$+ \frac{\rho}{2} \sum_{i=1}^{n} \big[\widehat{\mathrm{RGT}}_{\mathrm{tr},i}(\theta) - \varepsilon_r\big]_+^2 + \frac{\rho}{2} \sum_{i=1}^{n} \big[\widehat{\mathrm{IRV}}_{\mathrm{tr},i}(\theta) - \varepsilon_{ir}\big]_+^2 + \frac{\rho}{2} \big[\widehat{\mathrm{FR}}_{\mathrm{tr}}(\theta) - \varepsilon_f\big]_+^2. \tag{13}$$

We then perform alternating updates: (a) approximately minimize $\mathcal{L}_\rho$ over θ using stochastic gradient descent on mini-batches; (b) ascend in the dual variables (with projection onto $\mathbb{R}_+$) using, for example,

$$\lambda_i \leftarrow \Big[\lambda_i + \alpha_\lambda \big(\widehat{\mathrm{RGT}}_{\mathrm{mb},i}(\theta) - \varepsilon_r\big)\Big]_+, \quad \mu_i \leftarrow \Big[\mu_i + \alpha_\mu \big(\widehat{\mathrm{IRV}}_{\mathrm{mb},i}(\theta) - \varepsilon_{ir}\big)\Big]_+, \quad \nu \leftarrow \Big[\nu + \alpha_\nu \big(\widehat{\mathrm{FR}}_{\mathrm{mb}}(\theta) - \varepsilon_f\big)\Big]$$

where "mb" denotes mini-batch estimates. Economically, the multipliers act as endogenous "shadow prices" on violations: when regret or unfairness rises, training reweights the objective toward correcting it. Practically, the quadratic hinge terms stabilize training relative to a pure Lagrangian, which can oscillate when constraints are near-binding.

Numerical best responses as an adversarial inner loop. To compute $\phi_{i,\theta}(v, x)$ and its gradients, we require a best-response oracle. Because the mechanism is differentiable in bids by construction, we implement the inner maximization via projected gradient ascent in $b_i \in [0, 1]^m$. For each (v, x) in a mini-batch and each bidder i (or a randomly chosen subset of bidders to reduce cost), we solve

$$\max_{b_i \in [0,1]^m} \; u_i\big((b_i, v_{-i}); x, v_i\big)$$

approximately using K steps of a first-order method:

$$b_i^{(k+1)} \leftarrow \Pi_{[0,1]^m}\Big(b_i^{(k)} + \eta_{\mathrm{br}} \nabla_{b_i} u_i\big((b_i^{(k)}, v_{-i}); x, v_i\big)\Big),$$

possibly with momentum or Adam-style preconditioning. We typically use multiple random restarts (including the truthful report $b_i = v_i$) and keep the best iterate. Two implementation details matter for auditability. First, we log the achieved inner objective values and gradient norms; large residual gradients are a clear signal that the oracle is not close to optimal. Second, we explicitly track the number of steps K, restarts, and step sizes, since these choices determine an *oracle error* η that must later be reflected in the final regret certificate (as formalized by the oracle-gap logic in Proposition 5).

Differentiating through the max operator. If we could compute an exact maximizer $b_i^\star(v, x; \theta)$, then by Danskin-type arguments the gradient

of $\phi_{i,\theta}(v, x)$ with respect to θ can be taken by holding $b_i^\star$ fixed and differentiating the utility difference at that maximizer. With approximate best responses, we adopt the same heuristic—treat the final iterate $\tilde{b}_i$ as fixed when differentiating—which is standard in adversarial training. This choice is not merely computational convenience: it aligns with what we can later certify. The certificate will upper-bound true regret in terms of (i) the empirically estimated regret using a *tighter* adversary on holdout data, plus (ii) an explicit slack for the residual oracle error. In other words, we do not need training-time gradients to be exact; we need the evaluation-time adversary to be strong and its remaining error to be measurable and conservatively incorporated.

Practical approximations and their governance meaning. Three approximations are common and, in our view, should be treated as part of the mechanism's documented training protocol.

First, we may compute regret on a *subset* of bidders per mini-batch. This reduces cost from $O(n)$ best-response solves per sample to a constant number, but it introduces variance and can miss worst-case agents when n is large. A conservative deployment process therefore increases adversary strength at certification time (Section 6) and explicitly reports the maximum regret over all bidders on holdout.

Second, for fairness penalties that require comparing allocations across contexts (e.g., total-variation constraints over user pairs), we estimate $\widehat{\mathrm{FR}}_{\mathrm{tr}}(\theta)$ using within-batch pairings or a sampled neighborhood of contexts. This makes the training signal tractable, but it can underrepresent rare groups or long-tail contexts. Again, the remedy is not to claim training solves fairness, but to design the certification to target those tails (e.g., stratified holdout evaluation and stress tests).

Third, we often tighten training tolerances relative to the eventual policy tolerances, using $(\varepsilon_r^{\mathrm{tr}}, \varepsilon_{ir}^{\mathrm{tr}}, \varepsilon_f^{\mathrm{tr}})$ that are smaller than the desired deployment bounds. This "train conservatively" heuristic is economically interpretable: it builds a safety margin for generalization slack and for distribution drift. The cost, of course, is revenue; our framework makes that tradeoff explicit, and the certificate later quantifies whether the margin was sufficient.

Limitations. Constrained ERM with adversarial best responses is computationally intensive and nonconvex. We do not claim global optimality of the trained $\hat{\theta}$, nor do we claim that the training constraints alone guarantee safe deployment. Rather, training is the stage at which we *search* for a high-revenue mechanism that appears empirically compliant under a reasonably strong adversary. The next stage—certification—is where we insist on out-of-sample, high-confidence bounds under a tightened adversary and explicit diagnostics for the failure modes that training can mask.

6 Certification protocol: held-out evaluation with a tightened adversary

Training is inherently exploratory: we tune θ, adjust optimization hyperparameters, and (often) select a checkpoint. For an auditor, however, what matters is not whether constraints were *encouraged* during training, but whether the deployed mechanism satisfies incentive, IR, and fairness requirements *out of sample* with an explicit confidence level. We therefore separate *learning* from *certification*. The certification protocol takes a fixed candidate mechanism $\hat{\theta}$ and produces (i) empirical estimates of regret, IR violation, and fairness violation on data not used for fitting, and (ii) a conservative translation of these estimates into population-level bounds that hold with probability at least $1 - \delta$.

Holdout discipline and non-adaptivity. Let $\mathcal{S}_{\mathrm{ho}} = \{(v^{(s)}, x^{(s)})\}_{s=1}^{N_{\mathrm{ho}}}$ be a holdout sample drawn independently of training (or separated in time when data are mixing, using blocking). Certification is performed *once* on $\mathcal{S}_{\mathrm{ho}}$ after we commit to $\hat{\theta}$. To preserve the meaning of a $1 - \delta$ certificate, we treat all design choices that could respond to holdout outcomes—including which checkpoint to deploy, how many adversary steps to run, and which fairness subgroups to emphasize—as part of a pre-registered evaluation plan. When operational constraints force repeated certification attempts (e.g., iterative model updates), we account for this adaptivity by allocating failure probability across attempts (e.g., a Bonferroni-style split of δ) or by maintaining a fresh holdout stream.

Tightened adversary for holdout regret. Regret is the most delicate object to certify because it contains an inner maximization over misreports. The central principle is that the holdout adversary should be *at least as strong* as the adversary used in training. Concretely, for each $(v, x) \in \mathcal{S}_{\mathrm{ho}}$ and bidder i, we define a holdout regret integrand using a strengthened best-response oracle:

$$\tilde{\phi}_{i,\hat{\theta}}(v, x) \;=\; \max_{b_i \in \mathcal{A}_{\mathrm{ho}}(v,x)} \; u_i\big((b_i, v_{-i}); x, v_i\big) \;-\; u_i\big((v_i, v_{-i}); x, v_i\big),$$

where $\mathcal{A}_{\mathrm{ho}}(v, x) \subseteq [0, 1]^m$ is the feasible report set explored by the oracle (in the simplest case $\mathcal{A}_{\mathrm{ho}}(v, x) = [0, 1]^m$). The oracle itself is implemented by a higher-budget projected ascent routine: more steps K_{ho}, more random restarts R_{ho}, and (when useful) a small portfolio of step sizes and initializations that include $b_i = v_i$ and boundary points. We then set

$$\widehat{\mathrm{RGT}}_{\mathrm{ho},i}(\hat{\theta}) \;=\; \frac{1}{N_{\mathrm{ho}}} \sum_{(v,x) \in \mathcal{S}_{\mathrm{ho}}} \tilde{\phi}_{i,\hat{\theta}}(v, x), \qquad \widehat{\mathrm{RGT}}_{\mathrm{ho}}^{\max}(\hat{\theta}) \;=\; \max_{i \in [n]} \widehat{\mathrm{RGT}}_{\mathrm{ho},i}(\hat{\theta}).$$

Because the oracle is numerical, we also attach an explicit *oracle error* term $\eta_{\mathrm{ho}} \geq 0$ that upper-bounds how much the computed value could underestimate the true inner maximum. In practice, we make η_{ho} conservative by combining (i) observed improvement over the last iterations (a "no further progress" criterion), (ii) multiple restarts (so that the best found value is difficult to improve upon), and (iii) a stationarity diagnostic such as projected gradient norms. The certificate reports the full oracle configuration $(K_{\mathrm{ho}}, R_{\mathrm{ho}}, \eta_{\mathrm{ho}})$ so that the evaluation is reproducible.

Holdout IR and fairness estimates. IR violations do not require adversarial search. On holdout we compute

$$\widehat{\mathrm{IRV}}_{\mathrm{ho},i}(\hat{\theta}) \;=\; \frac{1}{N_{\mathrm{ho}}} \sum_{(v,x)\in\mathcal{S}_{\mathrm{ho}}} \max\Big\{0,\; -u_i((v_i, v_{-i}); x, v_i)\Big\}, \qquad \widehat{\mathrm{IRV}}_{\mathrm{ho}}^{\max}(\hat{\theta}) \;=\; \max_{i\in[n]} \widehat{\mathrm{IRV}}_{\mathrm{ho},i}(\hat{\theta}).$$

For fairness we compute a holdout analogue

$$\widehat{\mathrm{FR}}_{\mathrm{ho}}(\hat{\theta}) \;=\; \frac{1}{N_{\mathrm{ho}}} \sum_{(v,x)\in\mathcal{S}_{\mathrm{ho}}} \psi_{\hat{\theta}}(v, x),$$

where $\psi_{\hat{\theta}}$ is the per-sample (or per-batch) fairness-violation statistic. Because fairness constraints are often most binding in the long tail, we augment the global average with stratified reports: conditional estimates over protected groups, high-leverage contexts, or slices defined by policy-relevant features of x. When the fairness notion involves pairwise comparisons across contexts, we standardize the sampling scheme used to form pairs on holdout (e.g., fixed neighborhood sampling) to ensure that $\widehat{\mathrm{FR}}_{\mathrm{ho}}$ is a well-defined estimator of the target functional.

From holdout means to high-probability population certificates. The output of certification is not just a point estimate but a bound that holds with confidence $1 - \delta$. We therefore translate holdout estimates into population-level guarantees using slack terms $\Delta_r(N_{\mathrm{ho}}, \delta)$, $\Delta_{ir}(N_{\mathrm{ho}}, \delta)$, and $\Delta_f(N_{\mathrm{ho}}, \delta)$ that depend on sample size, failure probability, and the complexity of the hypothesis class. Formally, the certificate takes the form

$$\max_{i\in[n]} \mathrm{RGT}_i(\hat{\theta}) \;\leq\; \widehat{\mathrm{RGT}}_{\mathrm{ho}}^{\max}(\hat{\theta}) \;+\; \Delta_r(N_{\mathrm{ho}}, \delta_r) \;+\; \eta_{\mathrm{ho}},$$

$$\max_{i\in[n]} \mathrm{IRV}_i(\hat{\theta}) \;\leq\; \widehat{\mathrm{IRV}}_{\mathrm{ho}}^{\max}(\hat{\theta}) + \Delta_{ir}(N_{\mathrm{ho}}, \delta_{ir}), \qquad \mathrm{FR}(\hat{\theta}) \;\leq\; \widehat{\mathrm{FR}}_{\mathrm{ho}}(\hat{\theta}) + \Delta_f(N_{\mathrm{ho}}, \delta_f),$$

with $\delta_r + \delta_{ir} + \delta_f \leq \delta$ (for example, a simple equal split). Economically, the bound decomposes into three transparent components: what we observed on unseen data, what we pay for statistical uncertainty, and what we pay

for numerical under-optimization of the adversary. This decomposition matters in practice: when a certificate fails, it is diagnostically important to know whether failure is driven by observed violations, insufficient data, or an underpowered regret oracle.

Stress tests and "known unknowns." Holdout evaluation targets the distribution $\mathcal{D}$ represented by available samples. Deployment risk often comes from nearby but different environments: composition shifts in bidders, new contexts, or strategic adaptation over time. We therefore add an explicit stress-test layer that is not used to compute the main $1-\delta$ certificate but is reported alongside it. Examples include: (i) worst-slice evaluation over rare or high-impact contexts, (ii) perturbation tests that slightly modify v or x within plausible ranges to detect brittle discontinuities, and (iii) stronger adversaries that expand $\mathcal{A}_{\mathrm{ho}}(v,x)$ or increase K_{ho} beyond the pre-registered level. These stress tests are not formal guarantees; rather, they are governance tools that help decide whether the certified mechanism is robust enough for high-stakes deployment.

Failure modes and diagnostics. We treat certification as an opportunity to falsify overly optimistic conclusions. Four failure modes recur.

(1) Oracle weakness. If regret appears small only because the adversary is weak, then increasing K_{ho} or restarts should materially increase $\widehat{\mathrm{RGT}}_{\mathrm{ho}}^{\max}(\hat{\theta})$. We therefore report regret as a function of adversary budget and include convergence diagnostics (e.g., improvement curves and projected gradient norms).

(2) Selection bias from checkpointing. Choosing $\hat{\theta}$ by scanning many checkpoints against holdout performance invalidates nominal confidence levels. Our protocol avoids this by selecting $\hat{\theta}$ using training-only criteria (or a separate validation set) and using the holdout strictly once for certification.

(3) Tail risk masked by averages. Since $\mathrm{RGT}_i(\theta)$ and $\mathrm{FR}(\theta)$ are expectations, a small mean can coexist with rare but severe violations. We therefore report distributional diagnostics: quantiles of $\tilde{\phi}_{i,\hat{\theta}}(v,x)$, the maximum observed per-sample regret on holdout, and slice-wise maxima for fairness and IR.

(4) Temporal dependence and drift. When data are mixing rather than i.i.d., naive random splits can leak dependence. We mitigate this by block splitting and by periodic recertification on recent data, reporting certificate deterioration as an early warning signal.

Certification output. The final artifact is a short certificate report containing: the mechanism description and hash of $\hat{\theta}$; the holdout estimators $\widehat{\mathrm{RGT}}_{\mathrm{ho}}^{\max}$, $\widehat{\mathrm{IRV}}_{\mathrm{ho}}^{\max}$, and $\widehat{\mathrm{FR}}_{\mathrm{ho}}$; the computed slack terms $\Delta_r, \Delta_{ir}, \Delta_f$ and the oracle error η_{ho}; and the resulting population upper bounds. This is

the object that can be handed to a regulator or internal risk committee: it is explicit about uncertainty, explicit about numerical approximations, and explicit about where (and why) the guarantee could fail.

7 Main theory: uniform convergence and certified constrained optimality

We now make precise the statistical argument that turns finite-sample evaluations of regret, IR violation, and fairness violation into out-of-sample guarantees. The central difficulty is that incentive compatibility is assessed through an *inner maximization* (a best response), which can in principle amplify estimation error and destroy uniform convergence. Our approach is to (i) control the complexity of the *entire* integrand class induced by mechanisms in Θ, and (ii) exploit Lipschitzness and boundedness to ensure the inner maximization is stable.

Function classes for certification. Fix any bidder $i \in [n]$ and mechanism $\theta \in \Theta$. Define the per-sample regret integrand

$$\phi_{i,\theta}(v, x) \;:=\; \max_{v_i' \in [0,1]^m} u_i\big((v_i', v_{-i}); x, v_i\big) \;-\; u_i\big((v_i, v_{-i}); x, v_i\big),$$

as well as the per-sample IR-violation integrand

$$\zeta_{i,\theta}(v, x) \;:=\; \max\Big\{0,\; -u_i\big((v_i, v_{-i}); x, v_i\big)\Big\}.$$

For fairness we write $\psi_\theta(v, x)$ for the per-sample fairness-violation statistic whose expectation equals $\mathrm{FR}(\theta)$. We then consider the induced function classes

$$\mathcal{F}_r = \big\{\phi_{i,\theta} : \theta \in \Theta,\, i \in [n]\big\}, \quad \mathcal{F}_{ir} = \big\{\zeta_{i,\theta} : \theta \in \Theta,\, i \in [n]\big\}, \quad \mathcal{F}_f = \big\{\psi_\theta : \theta \in \Theta\big\}.$$

Uniform convergence for these classes yields simultaneous control of all bidders and all mechanisms that might plausibly be output by training or model selection (provided the evaluation protocol remains non-adaptive with respect to the holdout sample).

Boundedness and stability of the inner maximization. The regret integrand is a max over a continuum of misreports, so we first argue it remains well-behaved under our architectural assumptions. Because valuations lie in $[0,1]^m$, allocations satisfy $p_{i\ell} \in [0,1]$ with $\sum_i p_{i\ell} \le 1$, and payments are uniformly bounded by $t_i \in [0, B]$, utilities satisfy

$$-B \;\le\; u_i(b; x, v_i) \;\le\; m \qquad \Rightarrow \qquad 0 \;\le\; \phi_{i,\theta}(v, x) \;\le\; m + B.$$

More importantly, Lipschitzness of (p_θ, t_θ) in v implies that $\phi_{i,\theta}$ cannot oscillate sharply with small perturbations in types. Intuitively, if we perturb (v, x) slightly, then (a) the truthful utility changes smoothly, and (b) the best attainable utility under misreporting also changes smoothly because the maximization is taken over a compact set and the objective is itself Lipschitz. Formally, under mild regularity (e.g., existence of maximizers and an envelope argument), the mapping $(v, x) \mapsto \max_{v_i'} u_i((v_i', v_{-i}); x, v_i)$ inherits its Lipschitzness with a constant proportional to the Lipschitz constant of (p_θ, t_θ) and the bounded domain $[0, 1]^m$. The same reasoning applies to $\zeta_{i,\theta}$ (a hinge of a Lipschitz function) and to ψ_θ whenever fairness is defined as a bounded Lipschitz functional of allocations.

Uniform convergence via Rademacher complexity. Let $\mathcal{S} = \{(v^{(s)}, x^{(s)})\}_{s=1}^N$ be an evaluation sample drawn i.i.d. from $\mathcal{D}$ (extensions to mixing appear in Section 8). For any bounded function class $\mathcal{F}$, standard symmetrization yields, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \Big| \frac{1}{N} \sum_{s=1}^{N} f(v^{(s)}, x^{(s)}) - \mathbb{E}[f(v, x)] \Big| \;\le\; 2\,\mathfrak{R}_N(\mathcal{F}) \;+\; C\sqrt{\frac{\log(1/\delta)}{N}},$$

for a constant C depending only on the uniform bound on $\mathcal{F}$. Applying this bound to $\mathcal{F}_r, \mathcal{F}_{ir}, \mathcal{F}_f$ and union bounding over bidders where needed, we obtain slack terms $\Delta_r(N, \delta), \Delta_{ir}(N, \delta), \Delta_f(N, \delta)$ of the generic form

$$\Delta_\cdot(N, \delta) \;=\; O\left(\mathfrak{R}_N(\mathcal{F}_\cdot) + \sqrt{\frac{\log(1/\delta)}{N}} \right),$$

with $\log n$ factors when taking $\max_i$ and with constants that scale with $(m + B)$ for regret and with the corresponding uniform bounds for IR and fairness. Permutation equivariance helps here because it reduces effective complexity: intuitively, the class cannot represent bidder-specific idiosyncrasies, which tightens $\mathfrak{R}_N(\mathcal{F}_r)$ relative to an unconstrained architecture.

PAC-Bayes as a data-dependent alternative. Worst-case complexity bounds can be pessimistic for over-parameterized neural mechanisms. A complementary route is PAC-Bayes, which controls the *posterior-averaged* generalization gap in terms of a KL divergence to a prior. Concretely, let P be a prior over parameters θ (e.g., a Gaussian around an initialization), and let Q be a posterior distribution produced by training (for instance, a distribution over nearby checkpoints or a Gaussian around $\hat{\theta}$ with learned scale). For any bounded loss $f_\theta \in [0, U]$, a standard PAC-Bayes inequality implies that with probability at least $1 - \delta$ over $\mathcal{S}$,

$$\mathbb{E}_{\theta \sim Q}\big[\mathbb{E}[f_\theta] - \widehat{\mathbb{E}}[f_\theta]\big] \;\le\; O\left(\sqrt{\frac{\mathrm{KL}(Q \| P) + \log(1/\delta)}{N}} \right),$$

where $\widehat{\mathbb{E}}$ denotes the empirical mean on $\mathcal{S}$. Taking f_θ to be $\phi_{i,\theta}$ (or ψ_θ) yields a certificate for the randomized mechanism induced by Q, and, via standard arguments (e.g., choosing Q concentrated near $\hat{\theta}$), also provides a practical, data-dependent upper bound for the deterministic candidate. The policy-relevant point is that PAC-Bayes can turn "large network" into "small effective description length" when training finds a stable, low-complexity solution.

Accounting for approximate best responses. Because $\phi_{i,\theta}$ contains an inner maximum, certification must acknowledge that we compute best responses numerically. Suppose the regret oracle returns η-approximate best responses in the sense that for each (v, x) it outputs a value $\tilde{\phi}_{i,\theta}(v, x)$ satisfying

$$\tilde{\phi}_{i,\theta}(v, x) \;\geq\; \phi_{i,\theta}(v, x) - \eta.$$

Then empirical regret computed using $\tilde{\phi}$ underestimates true empirical regret by at most η, and consequently any population certificate acquires an additive η term. This separation is economically important: it distinguishes statistical uncertainty (which decays with N) from computational underoptimization (which decays with adversary budget and algorithmic improvements).

Certified constrained near-optimality. We can now state the main guarantee in a form aligned with the constrained learning objective. Let $\hat{\theta}$ be an (approximate) solution to the empirical problem

$$\max_{\theta\in\Theta} \widehat{\mathrm{REV}}(\theta) \quad \text{s.t.} \quad \max_i \widehat{\mathrm{RGT}}_i(\theta) \leq \varepsilon_r,\ \max_i \widehat{\mathrm{IRV}}_i(\theta) \leq \varepsilon_{ir},\ \widehat{\mathrm{FR}}(\theta) \leq \varepsilon_f,$$

where empirical quantities are evaluated on an independent sample and where $\widehat{\mathrm{RGT}}_i$ is computed with an η-approximate oracle. Then, with probability at least $1 - \delta$,

$$\max_i \mathrm{RGT}_i(\hat{\theta}) \;\leq\; \max_i \widehat{\mathrm{RGT}}_i(\hat{\theta}) + \Delta_r(N, \delta_r) + \eta, \qquad \max_i \mathrm{IRV}_i(\hat{\theta}) \;\leq\; \max_i \widehat{\mathrm{IRV}}_i(\hat{\theta}) + \Delta_{ir}(N, \delta_{ir}),$$

$$\mathrm{FR}(\hat{\theta}) \;\leq\; \widehat{\mathrm{FR}}(\hat{\theta}) \;+\; \Delta_f(N, \delta_f), \qquad \delta_r + \delta_{ir} + \delta_f \leq \delta.$$

Moreover, letting $\Theta^\star = \{\theta \in \Theta : \mathrm{RGT}(\theta) \leq 0,\ \mathrm{IRV}(\theta) \leq 0,\ \mathrm{FR}(\theta) \leq 0\}$ denote the (unknown) population-feasible set, the expected revenue satisfies the near-optimality bound

$$\mathrm{REV}(\hat{\theta}) \;\geq\; \sup_{\theta\in\Theta^\star} \mathrm{REV}(\theta) \;-\; O(\varepsilon_r + \varepsilon_{ir} + \varepsilon_f + \Delta_r + \Delta_{ir} + \Delta_f + \eta)\,,$$

where the big-O hides only universal constants and the uniform bounds on utilities and payments.

Interpretation and limitations. The theorem makes a specific tradeoff explicit. We can only certify what our class Θ can express and what our data can justify: tighter constraint tolerances reduce the feasible set and may lower revenue, while richer mechanisms increase revenue potential but worsen statistical slack through $\mathfrak{R}_N(\mathcal{F})$. This tension is not merely mathematical. In practice, regulators care about *documented* upper bounds on violations, and those bounds meaningfully depend on (i) architectural choices that control Lipschitzness and symmetry, (ii) the size and representativeness of evaluation data, and (iii) the strength and transparency of the regret oracle. Finally, our guarantee is relative to the best mechanism *within* Θ that satisfies the exact constraints; if the true economic optimum lies outside Θ (e.g., because of unmodeled bidder heterogeneity or richer type spaces), the certificate remains valid but the revenue comparison is necessarily class-conditional.

8 Extensions

We have stated the main theory under an i.i.d. evaluation sample primarily to keep the logic transparent. In many deployments, however, observations arrive sequentially and are neither independent nor identically distributed: bidder populations drift, contexts exhibit seasonality, and platform-side constraints induce temporal correlation in outcomes. We also often face richer private information than a pure valuation vector (e.g., budgets), and we may wish to certify mechanisms that operate across time rather than in a single static snapshot. Finally, it is natural to ask when classical closed-form mechanisms can be incorporated to strengthen guarantees rather than replaced by a black-box network. We briefly sketch how each of these extensions fits into the same certification template.

Weak dependence and mixing over time. Suppose we observe a time series $\{(v^{(s)}, x^{(s)})\}_{s=1}^{N}$ generated by a stationary process with weak dependence, rather than i.i.d. draws. A standard assumption in learning with dependent data is β-mixing (absolute regularity), with coefficients $\beta(t)$ that decay as the lag t grows. Intuitively, when $\sum_{t\geq 1}\beta(t)$ is small (or $\beta(t)$ decays geometrically), blocks of observations far apart behave approximately as if independent. Under such conditions, uniform convergence for bounded (and, in our case, Lipschitz) function classes continues to hold with an *effective sample size* smaller than N but still growing linearly in N up to dependence factors.

A convenient route is a blocking argument: partition $\{1,\dots,N\}$ into K blocks of length ℓ separated by gaps of length g, so that observations in different retained blocks are nearly independent. For a bounded class $\mathcal{F}$ with

envelope U, one obtains bounds of the schematic form

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{N}\sum_{s=1}^{N} f(v^{(s)},x^{(s)})-\mathbb{E}[f(v,x)]\right| \;\le\; O\!\left(\mathfrak{R}_K(\mathcal{F})+U\sqrt{\frac{\log(1/\delta)}{K}}\right)+O(U\,K\,\beta(g))\,,$$

where $K\approx N/(\ell+g)$. Choosing g so that $\beta(g)$ is small trades off bias (from dependence) against variance (from fewer effective blocks). The economic interpretation mirrors the i.i.d. case: statistical slack shrinks with data, but more slowly when the market is "sticky" over time. Importantly, the architectural levers emphasized above still matter. Lipschitzness and equivariance reduce the complexity terms (e.g., via Rademacher or covering-number bounds) regardless of whether the data are independent; dependence affects primarily the concentration step.

A practical implication for auditors is that the certificate should record not only N but also the dependence model used to justify concentration. When stationarity is questionable (e.g., known regime shifts), one can still certify *windowed* performance: apply the same bounds to rolling windows in which the process is plausibly stable, and treat re-certification as part of the governance process rather than a one-time event.

Private budgets and additional type components. Many applications, especially in advertising and procurement, impose budget constraints that are privately known or only imperfectly observable. A minimal extension is to augment bidder i's type from $v_i\in[0,1]^m$ to $\tau_i=(v_i,B_i)$, where $B_i\in[0,\overline{B}]$ is a budget cap. The mechanism now outputs $(p_\theta(\tau,x),t_\theta(\tau,x))$ with an additional feasibility constraint

$$0\le t_{\theta,i}(\tau,x)\le B_i \qquad \forall i.$$

This constraint can again be enforced by construction, for example by parameterizing an unconstrained payment $\tilde{t}_i\ge 0$ and setting $t_i=\min\{\tilde{t}_i,B_i\}$ (or a smooth approximation). Utilities remain quasi-linear up to the cap:

$$u_i(b;x,\tau_i)=\sum_{\ell=1}^{m} p_{i\ell}(b,x)\,v_{i\ell}-t_i(b,x),$$

and regret is defined exactly as before, except that deviations range over reports (v_i',B_i') in the allowed domain. Two modeling choices are common. If budgets are *verifiable* (the platform can enforce $t_i\le B_i$ and bidders cannot benefit from overstating), we restrict deviations to $B_i'\le B_i$; if budgets are *cheap talk*, we allow B_i' in the full interval and let incentives discipline truthful reporting. Either way, the certification logic is unchanged: the regret integrand is still a max of a bounded Lipschitz objective over a compact set, hence remains stable under the same regularity conditions, and uniform convergence applies to the induced function classes.

The main limitation is conceptual rather than statistical: budgets break the classic equivalence between DSIC and envelope-based payment formulas for additive valuations, so the mechanism class Θ matters more. In particular, a network that is expressive enough to exploit budgets may also be expressive enough to create subtle incentive issues unless Lipschitzness and symmetry are carefully controlled. Our framework is well-suited to this reality because it does not assume a closed-form characterization of truthfulness; it measures and certifies deviations directly, but it must pay for that flexibility through computation of best responses and through complexity control.

Dynamic settings and stagewise certificates. Platforms rarely run a single auction; they run a sequence. Let $t = 1, \dots, T$ index stages, and let $(v^{(t)}, x^{(t)})$ denote the types and context at stage t. A dynamic mechanism may couple allocations and payments across stages (e.g., pacing, carry-over budgets, or frequency capping). In such environments, full dynamic incentive compatibility is a statement about deviations in an entire *reporting policy* across time, which is typically intractable to certify without strong structural assumptions.

A pragmatic compromise is a *stagewise* certificate. We treat each stage as a (possibly context-augmented) static mechanism and certify that, conditional on the realized history $h^{(t)}$ that the mechanism observes, misreporting at stage t yields little gain relative to truthful reporting at that stage. Formally, define a per-stage regret integrand

$$\phi_{i,\theta}^{(t)}(v^{(t)}, x^{(t)}, h^{(t)}) = \max_{v_i' \in [0,1]^m} u_i^{(t)}\big((v_i', v_{-i}^{(t)}); x^{(t)}, h^{(t)}, v_i^{(t)}\big) - u_i^{(t)}\big((v_i^{(t)}, v_{-i}^{(t)}); x^{(t)}, h^{(t)}, v_i^{(t)}\big),$$

and certify $\mathbb{E}[\phi_{i,\theta}^{(t)}] \le \varepsilon_r^{(t)}$ uniformly over t (or on average). When the mechanism is myopic (no cross-stage coupling), this stagewise notion coincides with the static notion applied repeatedly. When coupling exists, stagewise regret becomes a bound on *one-step* deviations holding future play fixed; it is weaker than full dynamic truthfulness but can still be policy-relevant, especially when regulations are framed around per-auction transparency and when long-horizon deviations are practically limited by monitoring and budget constraints.

From a statistical viewpoint, stagewise certification is attractive: it reduces the deviation class from policies to per-stage reports, preserving compactness and enabling the same Lipschitz stability arguments. Dependence across t can be handled via the mixing extension above (or, when conditioning on histories, via martingale concentration under bounded differences). The cost is interpretability: the certificate must clearly state whether it bounds per-stage deviations, deviations over a restricted policy class, or full-horizon deviations, so that stakeholders understand the scope of the guarantee.

Embedding closed-form mechanisms as structure. Finally, we comment on when classical auction theory can strengthen guarantees. When the environment is close to a setting with a known optimal truthful mechanism (e.g., single-item with independent private values, or additive valuations with VCG-style structure), we can incorporate that structure into Θ rather than learning from scratch. There are two complementary benefits.

First, structure can reduce statistical complexity. If we parameterize allocations or payments through a low-dimensional family motivated by theory (e.g., monotone allocation rules with Myerson-style payment computation, or VCG with learned reserve adjustments), then $\mathfrak{R}_N(\mathcal{F}_\cdot)$ can be materially smaller than for a generic network, tightening $\Delta_r, \Delta_{ir}, \Delta_f$ and improving the practical sharpness of certificates. Second, structure can reduce the *oracle burden*. If parts of the mechanism are provably truthful by design, then regret computation becomes easier because the inner maximization is either identically zero in the structured limit or is constrained to a smaller deviation set arising from the learned residual components.

Concretely, one can view a hybrid mechanism as

$$(p_\theta, t_\theta) = \big(p_\alpha^{\text{base}}, t_\alpha^{\text{base}}\big) + \big(\Delta p_\beta, \Delta t_\beta\big),$$

where the base component is a closed-form truthful mechanism (or a differentiable approximation), and the residual is a Lipschitz, equivariant adjustment trained to meet fairness or business constraints. Certification then targets the combined mechanism, but we expect both regret and complexity to be smaller when the residual is small. The limitation is that theoretical structure is only as good as the model assumptions behind it; when independence, single-parameter structure, or quasilinearity fail, forcing a Myerson-style form can mis-specify incentives and reduce revenue. Our framework is designed to make this tradeoff explicit: adding structure may tighten certificates and improve robustness, but it can also restrict Θ and thereby lower the best achievable objective under the true environment.

Taken together, these extensions emphasize that certification is not tied to a single idealized statistical model. Rather, it is a modular pipeline: we enforce feasibility by construction, quantify incentive and fairness violations through empirically estimable functionals, and then choose the appropriate generalization theory (i.i.d., mixing, or martingale) and the appropriate type model (valuations, budgets, histories) to make the guarantees both mathematically defensible and operationally interpretable.

9 Experiments

We complement the theoretical certificate guarantees with experiments designed to answer three practical questions. First, in standard multi-item

benchmarks, do the architectural restrictions we advocate (feasibility layers, permutation equivariance, and Lipschitz control) materially change the revenue–incentives frontier relative to widely used learned-mechanism baselines? Second, in a context-rich advertising allocation problem with an explicit group fairness constraint, does certification remain informative at realistic sample sizes, or does the bound become vacuous? Third, in a procurement setting with volume discounts—where the economic primitives differ from "sell m items to n bidders"—does the same training-and-certification template remain useful, particularly when the best-response oracle is only approximate?

Common protocol and evaluation metrics. Across all environments we train mechanisms by constrained empirical risk minimization, maximizing empirical revenue subject to empirical constraints on (i) ex-post regret, (ii) ex-post individual-rationality violation, and (iii) the relevant fairness functional when present. We use a standard Lagrangian or augmented-Lagrangian procedure with multipliers updated on a held-out minibatch to reduce overfitting to the constraints. To compute empirical regret we employ a numerical best-response oracle: for each sampled (v, x) and bidder i, we approximately solve the inner maximization over misreports $v_i' \in [0,1]^m$ via projected gradient ascent on the differentiable surrogate utility induced by the mechanism network. This produces both a training-time estimate $\widehat{\mathrm{RGT}}_i(\theta)$ and, crucially for auditing, a *certification-time* estimate on an independent evaluation set. We report (a) empirical regret/IR/fairness on the training distribution, (b) the same quantities on the evaluation set, and (c) the corresponding certificate upper bounds (evaluation estimate plus the appropriate slack term), highlighting when certification is tight enough to be decision-relevant. In addition, we report the best-response oracle gap by running longer inner-loop optimization at evaluation time and measuring the increase in attained utility; this provides an empirical proxy for the η term in Proposition 5.

Baselines and ablations. We compare our mechanisms to four families that represent common design points in the recent literature. RegretNet is a generic, non-equivariant neural mechanism trained to minimize regret while maximizing revenue; it serves as a strong "black-box" baseline but typically lacks explicit symmetry and Lipschitz control. EquivarianceNet adds permutation-equivariant layers in the bidder dimension, aligning the architecture with exchangeability and often improving sample efficiency. RegretFormer represents a transformer-style architecture that can capture richer interactions among bidders and items, again usually without explicit Lipschitz calibration. ProportionNet is a fairness-oriented baseline that hard-codes proportional allocation heuristics (or their smooth relaxations) and

then learns payments; it tends to achieve low fairness violation but may sacrifice revenue or incentives because the allocation rule is heavily constrained. For our own method we ablate three components: (i) removing Lipschitz control (e.g., dropping spectral normalization or gradient penalties), (ii) removing equivariance while keeping feasibility and the same optimizer, and (iii) weakening the adversarial certification by replacing the best-response oracle with random misreports, which tests whether the certificate depends substantively on adversarial search rather than on incidental near-truthfulness.

(i) Standard multi-item multi-bidder benchmarks. We begin with synthetic benchmark distributions commonly used to evaluate learned mechanisms: additive valuations with $v_{i\ell}$ drawn independently from simple families (uniform, truncated normal, and mixtures that create "common high value" items), and contexts suppressed (x empty) to isolate the mechanism design problem. We vary (n, m) over a grid spanning small markets (e.g., $n \in \{2, 4\}$, $m \in \{2, 4\}$) to moderately sized markets (e.g., $n = 8$, $m = 8$) where the inner-loop regret computation becomes nontrivial. The principal outcome is a revenue–regret curve obtained by sweeping the target tolerance ε_r in training (and analogously ε_{ir} for IR). Two patterns are consistent across distributions. First, architectures that encode bidder symmetry reduce out-of-sample regret at a fixed revenue level: equivariant models dominate non-equivariant ones in the region where the regret constraint is tight, which is precisely the regime in which certification is economically meaningful. Second, explicit Lipschitz control reduces the variance of regret estimates across samples and narrows the train–test gap in regret and IR violation. In practical terms, when we remove Lipschitz control, we often observe mechanisms that appear nearly DSIC on the training sample but exhibit materially higher regret on the evaluation set; these are exactly the failures our certificate is intended to flag. The most salient comparison is that, at a fixed certified regret level, our approach achieves revenue comparable to (and often exceeding) RegretNet while producing substantially tighter certificates, suggesting that controlling sensitivity is not merely a theoretical convenience but improves the operational auditability of the learned mechanism.

(ii) Fairness-constrained ad allocation with contexts. We next consider a stylized display advertising allocation problem in which the context x encodes a user (or impression type) and eligibility constraints, and bidders represent advertisers partitioned into classes (e.g., protected categories, industries, or campaigns subject to policy constraints). Each bidder submits per-impression values $v_i(x)$ for a single slot (so $m = 1$) or a small set of slots/items (small m) representing multiple placements. We instantiate the fairness functional as a total-variation style constraint across users, in which the aggregated allocation probability for each advertiser class is required

to be similar across nearby users (as in the template $\mathrm{FR}(\theta)$ described earlier). We generate semi-synthetic values by combining observed covariates with randomized bidder-specific coefficients, allowing us to create realistic heterogeneity while retaining a known data-generating process.

The economic tension is transparent: stronger fairness constraints compress the ability to discriminate across contexts, lowering revenue, and they can also interact with incentives by changing the marginal effect of a bidder's report in different contexts. Empirically we find that fairness-constrained training produces mechanisms that satisfy fairness on the evaluation set, but only when the mechanism family is both equivariant (to reduce the effective complexity in the bidder dimension) and Lipschitz controlled (to prevent sharp context-dependent discontinuities that amplify estimation error in $\widehat{\mathrm{FR}}$). ProportionNet attains very low fairness violation but at a substantial revenue loss, and its incentive properties can be fragile because the restricted allocation rule forces payments to do too much of the work. RegretFormer can recover higher revenue, but without sensitivity control its fairness performance is less stable out of sample: in several configurations it meets the fairness tolerance on the training sample while violating it on the evaluation sample. Our certified approach, by contrast, tends to produce a "middle" outcome: revenue close to the best unconstrained models subject to certified fairness, with certificates that remain non-vacuous at moderate N. A key ablation result is that replacing adversarial misreports with random misreports can dramatically understate regret in this setting, because profitable deviations are structured and context-dependent; thus, the best-response oracle is not an implementation detail but part of what makes the evaluation credible.

(iii) Volume-discount procurement. Finally, we study a reverse-auction procurement environment motivated by platform purchasing and supply-chain contracting. Here the auctioneer purchases quantities of multiple goods, and suppliers have private cost information exhibiting volume discounts (equivalently, convexities in value in a selling formulation). We represent the procurement decision as allocating "demand units" across suppliers and items, with payments interpreted as transfers to suppliers; the feasibility constraint becomes a demand-fulfillment constraint and per-supplier capacity. Although the primitive differs from selling m items, the learned mechanism can be expressed in the same allocation–payment form by interpreting $p_{i\ell}$ as the probability (or fraction) of awarding unit ℓ to supplier i. We train to minimize expected procurement cost (negative revenue) subject to incentive and IR constraints; we also explore a fairness-like constraint corresponding to supplier diversification (penalizing excessive concentration in awards), which is economically motivated by resilience and antitrust considerations.

This setting stress-tests the regret oracle because suppliers' profitable deviations can be "global" across many units, creating a rugged optimization landscape. Consistent with Proposition 5, we observe that weaker inner-loop optimization leads to optimistic regret estimates; when we increase the oracle budget at certification time, estimated regret rises and the certificate correspondingly loosens. Importantly, the Lipschitz-controlled, equivariant architectures are more robust to this oracle gap: they exhibit smaller changes in regret as we strengthen the inner optimization, suggesting that smoothness reduces the prevalence of narrow, high-gain deviations that are hard to discover. From a governance perspective, this is an attractive property: it implies that the certificate is less sensitive to the precise numerical choices in the adversary, making the compliance story more stable.

Synthesis: what the experiments teach. Across the three domains, the empirical message aligns with the economic logic of our framework. When mechanisms are allowed to be highly sensitive and asymmetric, they can extract revenue in-sample but produce brittle incentives and fairness out of sample, rendering any a posteriori evaluation unreliable. Equivariance and Lipschitz control act as "regularizers with meaning": they constrain the mechanism in ways that correspond to exchangeability and continuity of responses to bids, improving both performance stability and the sharpness of certificates. Just as importantly, adversarial evaluation is essential: replacing best-response search with heuristic deviations can make constraint satisfaction appear easier than it is, particularly in context-rich environments where profitable deviations exploit interaction effects. We view these findings as a practical justification for treating certification as part of the mechanism design problem itself, rather than as an afterthought applied to a trained black box.

10 Discussion and Policy Implications

Our motivating premise is that a learned mechanism is not merely an optimization artifact but a decision rule that must be *defensible* to multiple audiences: engineers who deploy it, economists who reason about incentives, and auditors or regulators who need a verifiable account of what is guaranteed and at what confidence. The certificate framework we study provides a concrete object around which such accountability can be organized: an explicit upper bound on incentive violations (via regret), individual-rationality violations, and fairness violations, together with a stated failure probability. In this sense, certification plays the role that stress testing and capital buffers play in financial regulation: it translates complex model behavior into a small set of quantities that are legible for governance, while still being grounded in a rigorous out-of-sample statement.

How to report a certificate. A certificate should be reported as a tuple of *(i)* the empirical evaluation estimates on a holdout set, *(ii)* the corresponding slack terms (e.g., $\Delta_r(N,\delta)$, $\Delta_{ir}(N,\delta)$, $\Delta_f(N,\delta)$), *(iii)* the chosen confidence level $1-\delta$, and *(iv)* the numerical best-response oracle budget together with an explicit oracle-gap proxy. Concretely, for each bidder i we recommend reporting

$$\widehat{\mathrm{RGT}}_i(\hat{\theta}) \quad \text{and} \quad \widehat{\mathrm{RGT}}_i(\hat{\theta}) + \Delta_r(N,\delta) + \eta,$$

and analogously for IRV_i and FR. This is not a cosmetic choice: separating the estimate from the slack term makes clear whether non-compliance (if any) is due to observed violations or statistical uncertainty. Moreover, reporting the oracle budget and the empirical gap from a stronger adversary provides the practical counterpart of Proposition 5, preventing a common failure mode in which evaluation appears favorable simply because deviations were not adequately searched.

Interpreting certificate magnitudes. Because regret and fairness violations have different operational meanings, it is useful to complement scalar bounds with *economic calibration*. For regret, a bound of ε_r means that, in expectation under the reference distribution, a bidder can gain at most ε_r utility by misreporting. In markets where values are normalized to $[0,1]$, this quantity can be translated into an approximate "dollar" bound by rescaling with typical transaction sizes. For fairness, we recommend reporting both the expected violation $\mathrm{FR}(\hat{\theta})$ and salient conditional violations (e.g., worst-case over user groups in a finite audit set), even if the formal guarantee targets only the expectation. This mirrors common compliance practice: regulators often care about *tail risks* and subgroup harms, while the statistical theory typically controls an average. Being explicit about this distinction reduces the risk that a mathematically correct certificate is misinterpreted as a guarantee of uniform fairness across all contexts.

Monitoring and re-certification under distribution shift. The central limitation of any distributional certificate is that it is relative to a reference $\mathcal{D}$. In deployment, $\mathcal{D}$ can change due to seasonality, product changes, entry and exit of bidders, or deliberate strategic adaptation to the mechanism. A practical workflow therefore treats certification as a *living* requirement: we certify at launch and then periodically re-certify using fresh data, with an explicit monitoring layer in between. One operational approach is to maintain a rolling evaluation buffer and compute (i) drift statistics on (x,b) or estimated values, and (ii) online estimates of the certificate functionals. When drift is small, one can interpret the certificate as approximately valid; when drift exceeds a threshold, the mechanism is placed into a "heightened scrutiny" mode (e.g., conservative parameterization, tighter Lipschitz control, or fallback rules) until re-training and re-certification occur.

From a theoretical perspective, this suggests two complementary extensions. First, one can incorporate *robust* certificates that hold uniformly over a neighborhood of distributions around the empirical distribution (e.g., Wasserstein balls), trading off slack for resilience to modest shift. Second, one can analyze *mixing* or time-series regimes in which samples are not i.i.d.; while our global context allows mild temporal dependence, practical monitoring benefits from explicit bounds that account for effective sample size under mixing. Both directions connect naturally to existing tools in distributionally robust optimization and learning under dependence, but remain underdeveloped for multi-agent incentive constraints.

Compliance workflows: roles, logs, and change management. In organizational terms, we view certification as enabling a clear separation of responsibilities. The mechanism designer (or ML team) produces the trained parameters $\hat{\theta}$ and a "certificate report" containing the quantities above; the auditor (internal or external) verifies the report by re-running evaluation scripts on a locked holdout set and, critically, by re-running the best-response oracle with a prescribed minimum budget. This suggests simple but important engineering controls: immutable logs of mechanism outputs (p, t), versioning of θ and the oracle code, and retention of evaluation datasets used for certificates. Change management should treat any modification to the hypothesis class Θ (architecture, equivariance choices, Lipschitz enforcement), or to the fairness functional itself, as a *material* change requiring re-certification. In regulated settings, the appropriate analogy is a model-risk governance program: the certificate becomes a standardized artifact that can be reviewed, archived, and compared across model versions.

Policy relevance: what regulators can ask for. A regulator or platform policy team cannot be expected to inspect neural mechanism internals, but it can require a small set of standardized disclosures. We propose three. First, a statement of the targeted constraint semantics: regret as an approximation to DSIC, the precise IR notion (ex-post versus interim), and the fairness definition including the user distance d_c and class partition. Second, quantitative certificates with a specified δ (e.g., 5% or 1%), including a transparent accounting of the oracle gap. Third, an operational monitoring plan describing how drift is detected, how often re-certification occurs, and what fallback rule is used when certificates become non-informative. Importantly, such disclosures are compatible with business confidentiality: they need not reveal bidders' data or the mechanism weights, only performance bounds and procedures. This is a pragmatic virtue of certificate-based governance relative to more intrusive forms of oversight.

Limitations. Several limitations deserve emphasis. (i) *Regret is a proxy:* small ex-post regret implies approximate incentive compatibility but does not fully characterize equilibrium behavior in repeated or information-rich environments; bidders may learn and coordinate, and small one-shot deviations need not preclude profitable multi-period strategies. (ii) *Fairness is contestable:* no single functional captures all normative concerns, and a bound on one fairness metric can coexist with harms measured by another. Our framework is agnostic to the choice of FR, but the burden shifts to the policy process that selects it. (iii) *Computational scalability:* the inner maximization that defines regret is costly in high-dimensional type spaces, and approximate oracles can be brittle. Lipschitz control helps, but does not eliminate the risk that the adversary misses structured deviations. (iv) *Endogeneity and strategic distribution shift:* once deployed, the mechanism can change the distribution of observed bids and contexts (and, in ad markets, user composition), so the data used for re-certification may itself be policy-induced. This complicates both statistical inference and causal interpretation.

Open problems. These limitations point to several research directions. One is to develop *stronger and more interpretable* certificates, for example converting regret bounds into approximate payment identity statements or monotonicity diagnostics that economists find more transparent. Another is to integrate *online learning* with continual certification, providing guarantees that hold uniformly over time while the mechanism adapts. A third is to couple certification with *privacy* constraints, since many markets require that training and auditing respect bidder confidentiality; the interaction between differential privacy and incentive guarantees remains subtle. Finally, there is the broader question of *market-level* outcomes: revenue and constraint satisfaction are mechanism-level objects, but policy often cares about allocative efficiency, entry, and long-run welfare. Extending certificate-based design to these outcomes—particularly under strategic responses and platform feedback loops—is an important step toward making learning-based mechanism design a mature tool for high-stakes allocation.

Taken together, these considerations support a practical view of our contribution: certification is not a replacement for economic judgment, but a disciplined way to connect learning systems to the kinds of quantitative assurances that real-world governance demands. The next section concludes by summarizing what our framework guarantees today and what must be built to make such guarantees routine in deployed markets.

11 Conclusion

We set out from a simple observation: once allocation and pricing rules are learned from data, the central question is no longer only whether the mechanism performs well *on average*, but whether it can be accompanied by an intelligible and defensible account of what it guarantees. In classical mechanism design, incentive and participation properties are proved from closed-form structure (e.g., monotonicity and payment identities). In learned mechanism design, the mapping from reports and context to outcomes is typically too complex to inspect directly, and so the appropriate analog of a proof is an *out-of-sample certificate*: a quantitative bound on economically meaningful violations, together with an explicit confidence level and a transparent accounting of numerical approximation error.

Our framework formalizes this idea in a setting that is both economically standard and operationally relevant: multi-item allocation with additive values, context-dependent constraints, and a designer who chooses a parameterized mechanism from a structured hypothesis class. The central modeling move is to treat incentive compatibility, individual rationality, and fairness not as hard constraints that must hold pointwise, but as *population* constraints expressed through functionals—expected ex-post regret, expected IR violation, and an expected fairness violation measure. This choice is not merely for convenience. In markets such as advertising, cloud procurement, or sponsored recommendations, the objects that are actually monitored and audited are typically distributional summaries (averages over time, or over sampled user/campaign slices), and the operational goal is to ensure these summaries remain below tolerances that are meaningful at the scale of the business or policy regime.

Three structural ingredients make certification plausible. First, feasibility is handled *by construction*: architectural layers (softmax/Sinkhorn with optional slack capacity) enforce that allocations are valid for every input, eliminating an entire class of failure modes in deployment. Second, we impose symmetry and regularity through permutation equivariance and uniform Lipschitz control. Equivariance aligns the hypothesis class with the economic symmetry that bidders are interchangeable absent features, improving sample efficiency and reducing opportunities for idiosyncratic overfitting. Lipschitz control, while a blunt instrument, ensures that small changes in reports cannot produce arbitrarily large changes in allocations or payments, which stabilizes both learning and evaluation. Third, we make the role of computation explicit by incorporating a numerical best-response oracle for regret estimation, and by tracking how oracle approximation propagates to the final guarantees.

On top of these primitives, our main theoretical message is that constrained learning can be made *auditable*. Uniform convergence bounds for the regret, IR, and fairness function classes imply that held-out empirical es-

timates can be turned into population-level upper bounds with a computable slack. In particular, for a learned parameter $\hat{\theta}$, one can report a certificate of the form

$$\Big(\widehat{\mathrm{RGT}}_i(\hat{\theta}),\ \widehat{\mathrm{IRV}}_i(\hat{\theta}),\ \widehat{\mathrm{FR}}(\hat{\theta})\Big) \quad \text{and} \quad \Big(\widehat{\mathrm{RGT}}_i(\hat{\theta})+\Delta_r+\eta,\ \widehat{\mathrm{IRV}}_i(\hat{\theta})+\Delta_{ir},\ \widehat{\mathrm{FR}}(\hat{\theta})+\Delta_f\Big),$$

with an accompanying statement that, with probability at least $1-\delta$, the corresponding population violations are no larger than the reported upper bounds. This shifts the evaluation question from "do we believe the neural network is truthful?" to "do we believe the sampling assumptions, the held-out evaluation protocol, and the adversarial search budget?"—questions that can be answered with standard governance tools: data documentation, access control for holdout sets, and reproducible evaluation scripts.

The optimization side of the framework matches this logic. Training is posed as constrained empirical risk minimization: maximize estimated revenue subject to estimated constraints. Our main theorem states that if $\hat{\theta}$ approximately solves this empirical problem, then (with high probability) it satisfies the *population* constraints up to the slack terms, and it achieves near-optimal revenue relative to the best mechanism in the class that satisfies the target constraints. Economically, this near-optimality statement is the counterpart of what one expects from classical design: within a specified design space, we can approach the best revenue while respecting incentive, participation, and fairness limits. Statistically, it clarifies where the tradeoffs live: tighter tolerances, richer hypothesis classes, larger Lipschitz constants, higher-dimensional environments, and weaker mixing all push against the strength of the certificate, and thus against the designer's ability to claim compliance at a given confidence level.

The practical implication is that the learned-mechanism pipeline can be organized around *artifacts* that are legible outside the modeling team. A mechanism version can be shipped with: (i) a precise statement of the semantics of its constraints (what regret means, what IR means, what fairness notion is used), (ii) a quantitative certificate with a declared δ and a declared oracle budget, and (iii) an operational plan for monitoring and re-certification. In settings where regulators or platform policy teams demand accountability but cannot inspect source code or model weights, such artifacts offer a middle ground between unverifiable claims and intrusive disclosure. Moreover, the framework naturally supports internal "red teaming": the best-response oracle becomes a tool for systematic adversarial testing, analogous to security penetration tests, with the oracle-gap parameter η serving as a clean summary of how hard the adversary tried.

At the same time, we emphasize what this framework does *not* resolve. Ex-post regret is an economically meaningful proxy, but it is not a full theory of behavior in repeated, information-rich environments where bidders can condition on history, coordinate, or exploit feedback loops. Fairness metrics,

even when carefully chosen and transparently reported, are necessarily partial; a bound on one metric can coexist with harms expressed in another vocabulary. The Lipschitz and boundedness assumptions, while technically useful and often implementable through architecture and regularization, can exclude mechanisms that are desirable in theory, and can impose a real revenue cost in practice. Finally, certification remains distribution-relative: it certifies performance under $\mathcal{D}$ (or under a modeled dependence structure), and thus inherits the usual vulnerability of statistical guarantees to distribution shift—including shift that is itself induced by the deployed mechanism.

These limitations suggest a research agenda that is as much economic as it is statistical. One direction is to enrich the space of certifiable properties beyond regret-based proxies, for example by developing diagnostics for approximate monotonicity, approximate payment identities, or "approximate core" conditions in combinatorial settings, which may be more interpretable to economic stakeholders. A second direction is to develop certificates that are explicitly robust to shift, either by controlling violations uniformly over neighborhoods of distributions or by integrating online monitoring with sequential bounds that remain valid under continual adaptation. A third direction concerns computation: stronger best-response methods, tighter oracle-gap estimates, and more structured deviation classes could make regret evaluation both cheaper and more reliable in large-scale markets. A fourth direction is to connect mechanism-level certificates to market-level outcomes—efficiency, entry, and long-run welfare—especially in platform environments where the mechanism shapes participation and information.

We close with a broader perspective. Learning-based mechanism design is often presented as a way to sidestep analytic complexity by letting data and optimization "discover" good rules. Our view is that, for high-stakes allocation, discovery is not enough: what matters is discovery plus documentation. The certificate framework developed here is one step toward making that documentation principled. It does not replace economic judgment about which constraints are normatively appropriate or which tradeoffs are acceptable. Rather, it provides a disciplined language for stating what is being optimized, what is being bounded, what could go wrong, and how confident we are. If learned mechanisms are to become routine tools in regulated or reputationally sensitive markets, the ability to produce such statements—and to update them as markets evolve—will be at least as important as incremental gains in revenue.