LEARNING SAMPLES IMPORTANCE: PARAMETERIZING DUAL VARIABLES IN EVERYWHERE LEARNING.

Anonymous authors
Paper under double-blind review

ABSTRACT

Everywhere learning provides a principled framework for training AI models under constraints that must hold throughout the data distribution. In the dual domain, these pointwise constraints give rise to functional dual variables. In this work, we propose to learn these dual variables, motivated by the fact that their values encode useful information about the underlying constrained problem. By representing the dual variable as a parametric function of each sample, we enable the learned multiplier to be evaluated on new, unseen samples. This contrasts with standard empirical dual formulations, which assign an independent multiplier to each training sample. We characterize the error in the recovered primal solution induced by restricting the dual variable to a parametric function class and show that it is controlled by how well this class approximates the optimal statistical multiplier. Moreover, we show that the learned parametric multiplier retains the sensitivity interpretation of the optimal statistical multiplier, yielding approximate sensitivity guarantees that extend beyond the samples used for training. We empirically validate our theory across a variety of everywhere learning tasks, showing that the resulting constrained problems can be solved efficiently and that the learned dual variables provide meaningful representations of sample-level sensitivity.

1 INTRODUCTION

Constrained learning provides a natural framework for training AI systems subject to multiple requirements (Zafar et al., 2017; Cotter et al., 2019; Chamon & Ribeiro, 2021). Expressing requirements as constraints prevents improvements in one criterion from compensating for violations of another. Everywhere learning proposes to impose requirements pointwise over the data distribution instead of in expectation, preventing imbalances in performance across data regions (Boero et al., 2026; Hounie et al., 2026). Solving the resulting problem naturally leads to the dual domain, where the dual multipliers are themselves functions of the data. In classical optimization, optimal dual multipliers carry a sensitivity interpretation, quantifying how changes to individual constraints affect the optimal value (Boyd & Vandenberghe, 2004; Rockafellar, 1993; Shapiro et al., 2021). This motivates our central question of whether these dual multipliers can be learned.

(C1) We propose to learn dual multipliers by parameterizing them as a function of the samples.

Rather than assigning an independent multiplier to every training sample, we optimize over a parametric family Λ_Φ, allowing the learned multiplier to be evaluated on unseen samples. This, however, restricts the original functional dual domain to Λ_Φ, and it is therefore not immediate that the resulting dual problem still recovers a suitable primal solution.

(C2) We provide near-feasibility and near-optimality guarantees for the primal solution recovered from the parametric dual problem.

We show that the primal solution induced by the learned parametric dual remains approximately feasible and optimal. In particular, the additional error introduced by parameterizing the dual domain is controlled by how well Λ_Φ approximates an optimal multiplier. Our goal, however, is not only to recover a good primal model, but also to preserve the information carried by the multipliers themselves.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(C3) We show that the learned multiplier retains the sensitivity role of the optimal dual variable.

The learned multiplier continues to approximately quantify how perturbing individual constraints can affect the optimal value of the problem. Unlike primal recovery, extending this interpretation to unseen samples additionally requires the dual class to generalize, revealing a natural tradeoff: richer parameterizations can better approximate the optimal multiplier but may incur larger statistical error. We validate these results across four diverse machine-learning problems spanning image classification, agentic workflows, language-model unlearning, and optimal power flow.

(C4) We show empirically that parametric duals preserve primal performance while learning transferable sample-level sensitivity.

Across tasks, expressive dual parameterizations approach the performance of sample-wise multipliers, remain effective as the number of constraints grows, and accurately generalize multiplier information to unseen samples.

2 EVERYWHERE LEARNING

Let $\{\mathcal{D}_i\}_{i=0}^m$ be probability distributions over input-output pairs $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Given bounded loss functions $\ell_i : \mathcal{Y} \times \mathcal{Y} \to [0, B]$, for $i = 0, \ldots, m$, and constraint thresholds $c_i \in \mathbb{R}$, for $i = 1, \ldots, m$, the *everywhere learning problem* seeks the parametric predictor $f_\theta : \mathcal{X} \to \mathcal{Y}$ in the hypothesis class $\mathcal{H} := \{f_\theta : \theta \in \Theta \subseteq \mathbb{R}^p\}$ that solves the constrained problem:

$$\begin{aligned} f_\theta^* = \arg\min_{f_\theta \in \mathcal{H}} \quad & \mathbb{E}_{(x,y)\sim\mathcal{D}_0}\Big[\ell_0\big(f_\theta(x), y\big)\Big] \\ \text{s. to} \quad & \ell_i\big(f_\theta(x), y\big) \le c_i \quad \mathcal{D}_i\text{-a.e.} \end{aligned} \tag{P}$$

The everywhere learning problem P provides a formulation for learning predictors subject to multiple requirements. Expressing these requirements as constraints makes their prescribed tolerances explicit and prevents improvements in one criterion from compensating for violations of another, as may occur when minimizing weighted combinations of losses. Moreover, the almost-everywhere constraints require each requirement to hold throughout its corresponding data distribution. This prevents good performance in some regions from compensating for violations in others, as can occur when constraints are imposed only in expectation.

Henceforth, we consider a single constraint $m = 1$, set its tolerance $c = 0$ without loss of generality, denote $z = (x, y) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ be a compact representation of the input-output pairs, and define $\ell(f_\theta(x), y) = \ell(f_\theta, z)$, for notational convenience.

2.1 EVERYWHERE DUAL PROBLEM

A standard approach when solving constrained problems is to turn to the dual domain . Since Problem P imposes a constraint pointwise at every z, its dual variable $\lambda : \mathcal{Z} \to \mathbb{R}_+$ is itself a nonnegative function, where $\lambda(z)$ weights the constraint imposed at point z. We restrict these functions to the space of nonnegative functions integrable with respect to the constraint distribution, that is, $\lambda \in L_1^+(\mathcal{D})$. The Lagrangian function $\mathcal{L} : \mathcal{H} \times L_1^+(\mathcal{D}) \to \mathbb{R}$ aggregates the objective value with the constraint loss weighted pointwise by the corresponding dual variable,

$$\mathcal{L}(f_\theta, \lambda) := \mathbb{E}_{z\sim\mathcal{D}_0}\left[\ell_0(z, f_\theta)\right] + \mathbb{E}_{z\sim\mathcal{D}}\left[\lambda(z)\ell(z, f_\theta)\right]. \tag{1}$$

For a fixed multiplier λ, the Lagrangian function reduces the constrained problem to an unconstrained optimization over the predictor. The function that takes a multiplier and returns the Lagrangian minimization is denoted the *dual function*,

$$g(\lambda) := \mathcal{L}\big(f_\theta(\lambda), \lambda\big) = \min_{f_\theta \in \mathcal{H}} \mathcal{L}(f_\theta, \lambda). \tag{2}$$

The multiplier λ therefore determines how the pointwise constraints are weighted when optimizing the predictor. Ideally, we seek a weighting for which the resulting Lagrangian minimizer $f_\theta(\lambda)$ is feasible for Problem P while achieving a low objective value. A principled approach for searching

over such weightings is given by the dual problem, which maximizes the dual function over all admissible nonnegative multipliers,

$$\lambda^\star \in \arg \max_{\lambda \in L_1^+(\mathcal{D})} g(\lambda), \tag{D}$$

and yields $D^\star = g(\lambda^\star)$ as its optimal value. Previous work establishes the connection between Problem P and its dual formulation. For convex problems, the primal and dual optimal values coincide under mild conditions. For the nonconvex parametrizations commonly used in deep learning, this equivalence need not hold exactly; nevertheless, the resulting duality gap can be controlled by the richness of the predictor class Boero et al. (2026). Thus, the dual formulation provides a principled approach for studying and solving the everywhere learning problem.

Beyond its role in solving the constrained problem, the optimal multiplier $\lambda^\star(z)$ provides a sample-level measure of how strongly each constraint shapes the learned predictor. In the next section, we study the conditions under which the optimal multiplier $\lambda^*(z)$ can be learned from a parametric family $\lambda_\phi(z)$ while simultaneously solving the constrained problem from samples. This differs from the standard empirical dual formulation, which assigns an independent multiplier to each training sample and thus can not provide dual values for new samples Boero et al. (2026).

3 Parametric Empirical Dual Problem

As in standard machine learning settings, the distributions D_0 and D are unknown and can only be accessed through samples. We therefore replace the expectations in equation 1 by sample averages over datasets $S_0 = \{z_n\}_{n=1}^{N_0}$ and $S = \{z_n\}_{n=1}^{N}$, yielding the empirical Lagrangian

$$\hat{L}(f_\theta, \lambda) := \frac{1}{N_0} \sum_{z_n \in S_0} \ell_0(z_n, f_\theta) + \frac{1}{N} \sum_{z_n \in S} \lambda(z_n) \ell(z_n, f_\theta). \tag{3}$$

As in the statistical setting, fixing the dual multiplier λ and minimizing the empirical Lagrangian over the hypothesis class defines the empirical dual function

$$\hat{g}(\lambda) = \min_{f_\theta \in \mathcal{H}} \hat{L}(f_\theta, \lambda). \tag{4}$$

The corresponding empirical dual problem would maximize $\hat{g}(\lambda)$ over nonnegative multiplier functions. Direct optimization over the infinite-dimensional space $L_1^+(\mathcal{D})$, however, is generally intractable. To obtain a finite-dimensional formulation, we propose learning the dual multiplier within a bounded parametric hypothesis class Λ_Φ,

$$\Lambda_\Phi = \big\{ \lambda_\phi : \mathcal{Z} \to \mathbb{R}_+ \;\big|\; \lambda_\phi \in L_1^+(\mathcal{D}),\ \phi \in \Phi \subseteq \mathbb{R}^q, \|\lambda_\phi\|_\infty \leq \Gamma \big\}. \tag{5}$$

We then define the *empirical parametric dual problem* by maximizing the empirical dual function over Λ_Φ,

$$\hat{\lambda}_\Phi^\star \in \arg\max_{\lambda_\phi \in \Lambda_\Phi} \hat{g}(\lambda_\phi), \tag{$\widehat{\mathrm{D}}_\Phi$}$$

with optimal value $\hat{D}_\Phi^\star = \hat{g}(\hat{\lambda}_\Phi^\star)$.

This approach contrasts with the standard construction of an *empirical dual variable* $\hat{\lambda} \in \mathbb{R}_+^N$, in which one multiplier is assigned independently to each training sample. Such a representation is defined only on the observed dataset and therefore provides no mechanism for assigning multipliers to unseen samples. By contrast, each $\lambda_\phi \in \Lambda_\Phi$ defines a multiplier function over the entire domain $\mathcal{Z}$, allowing the learned dual variable to be evaluated beyond the training set.

We remark that replacing expectations by sample averages and restricting the dual domain to the parametric class Λ_Φ constitute two approximations required to obtain a tractable formulation of the statistical dual problem D. These modifications, however, imply that problem $\widehat{\mathrm{D}}_\Phi$ is no longer the dual of the original statistical primal problem P. In this section, we show that both this discrepancy and the error in interpreting $\hat{\lambda}_\Phi^\star$ as the sensitivity of the statistical problem vanish as the parametric family better approximates the optimal multiplier λ^* and the number of samples increases.

3.1 PRIMAL RECOVERY FOR $\widehat{\mathrm{D}}_\Phi$

The parametrized empirical dual problem equation $\widehat{\mathrm{D}}_\Phi$ introduces two approximations with respect to the statistical dual problem equation D. Therefore, to characterize the effect of these approximations, we require conditions that control each of them. The sampling error is governed by uniform convergence of the objective and constraint losses, while the dual-parameterization error is determined by how well Λ_Φ can approximate an optimal statistical multiplier. We formalize these two requirements next.

Assumption 1 (Uniform convergence)**.** *There exist functions $\zeta_0(N_0, \delta)$ and $\zeta(N, \delta)$, monotonically decreasing to zero in N_0 and N, respectively, such that, with probability at least $1 - \delta$,*

$$\sup_{f_\theta \in \mathcal{H}} \left| \mathbb{E}_{D_0}[\ell_0(f_\theta, z)] - \frac{1}{N_0} \sum_{z_n \in S_0} \ell_0(f_\theta, z_n) \right| \leq \zeta_0(N_0, \delta), \tag{6}$$

$$\sup_{f_\theta \in \mathcal{H}} \left| \mathbb{E}_{D}[\ell(f_\theta, z)] - \frac{1}{N} \sum_{z_n \in S} \ell(f_\theta, z_n) \right| \leq \zeta(N, \delta). \tag{7}$$

Assumption 2 (Approximation of the optimal multiplier)**.** *Let $\lambda^\star$ be an optimal multiplier of the statistical dual problem. The approximation error induced by Λ_Φ is*

$$\nu_\Phi := \inf_{\lambda_\phi \in \Lambda_\Phi} \|\lambda_\phi - \lambda^\star\|_{L_1(D)}. \tag{8}$$

Assumption 1 is a standard learnability condition controlling the discrepancy between population and empirical losses uniformly over $\mathcal{H}$. An example of obtaining those rates from controlled Rademacher complexity of $\mathcal{H}$ is discussed in Appendix A.1. Assumption 2, on the other hand, measures how well the parametric dual family can approximate an optimal statistical multiplier. The assumption itself is mild, as the approximation error is not required to be small a priori; rather, its value explicitly quantifies the price of restricting the functional dual domain to Λ_Φ.

Under these assumptions we can control the discrepancy between the optimal statistical and parametrized empirical dual values.

Lemma 1 (Dual learnability)**.** *Let Assumptions 1 and 2 hold. Recall that, by definition of Λ_Φ, every $\lambda_\phi \in \Lambda_\Phi$ satisfies $\|\lambda_\phi\|_\infty \leq \Gamma$. Then, with probability at least $1 - 3\delta$,*

$$\left| D^\star - \widehat{D}^\star_\Phi \right| \leq B\nu_\Phi + \zeta_0(N_0, \delta) + \Gamma\zeta(N, \delta) =: \Delta_D. \tag{9}$$

Proof. See Appendix A.2 □

Lemma 1 separates the dual approximation error into two components. The terms ζ_0 and $\Gamma\zeta$ arise from replacing population quantities by finite-sample estimates, whereas $B\nu_\Phi$ is the additional error introduced by restricting the dual domain to Λ_Φ. Closeness of the optimal dual values, however, does not by itself guarantee that a primal minimizer is an approximate solution of P. To recover such guarantees, we require extra assumption on the structure of the problem P, which ensures that it can be well approximated through the dual problem.

Assumption 3 (Convex losses and curvature)**.** *For every $y \in \mathcal{Y}$, the losses $\ell(\cdot, y)$ and $\ell_0(\cdot, y)$ are convex, and $\ell_0(\cdot, y)$ is μ-strongly convex in its prediction argument.* [1]

Assumption 4 (Lipschitz losses)**.** *For every $y \in \mathcal{Y}$, the losses $\ell(\cdot, y)$ and $\ell_0(\cdot, y)$ are L-Lipschitz continuous in their prediction argument, that is,*

$$|\ell_i(\hat{y}_1, y) - \ell_i(\hat{y}_2, y)| \leq L\|\hat{y}_1 - \hat{y}_2\|. \tag{10}$$

Assumption 5 (Approximate convexity of the model class)**.** *The hypothesis class $\mathcal{H}$ approximates its closed convex hull. Specifically, letting*

$$\overline{\mathcal{H}} := \overline{\mathrm{conv}}(\mathcal{H}), \tag{11}$$

for every $f \in \overline{\mathcal{H}}$ there exists $f_\theta \in \mathcal{H}$ such that

$$\|f_\theta - f\|_\infty \leq \nu_{\mathcal{H}}. \tag{12}$$

[1] For the standard-Lagrangian argument, the strong-convexity condition is understood in the prediction norm controlling the constraint losses. This requirement is avoided by the augmented Lagrangian used in our experiments, whose dual smoothness follows directly from augmentation; see Appendix A.5.

Assumption 6 (Strict Slater condition)**.** *There exists $\bar{f}_\theta \in \mathcal{H}$ and $s > 0$ such that*

$$\ell(\bar{f}_\theta, z) + L\nu_{\mathcal{H}} \leq -s, \qquad D\text{-a.e.} \tag{13}$$

Assumptions 3 and 4 are standard regularity conditions and are satisfied by common losses on bounded prediction domains, including the squared loss. Assumption 5 measures how closely the original hypothesis class can represent predictors in its convex hull. It holds with $\nu_{\mathcal{H}} = 0$ for convex hypothesis classes. For neural networks, finite combinations of predictors can be represented by increasing model width and combining the corresponding subnetworks at the output layer, consistent with the view of wide neural networks as linear combinations of learned features (Bengio et al., 2005). Thus, $\nu_{\mathcal{H}}$ can be interpreted as a capacity-dependent approximation error which can decrease as the model class becomes richer. Finally, Assumption 6 is a strengthened Slater condition. It requires the existence of a model which can satisfy the constraints with a margin $L\nu_{\mathcal{H}}$, and is a mild strengthening of requiring the existence of a feasible predictor.

Under these assumptions, the primal minimizer associated with the learned parametric multiplier is both approximately feasible and approximately optimal for the original statistical problem P.

Theorem 1 (Statistical primal recovery)**.** *Let Assumptions 1–6 hold, and let $\widehat{f}_{\theta_\Phi}$ be a primal minimizer associated with the optimal parametrized empirical multiplier,*

$$\widehat{f}_{\theta_\Phi} \in \arg\min_{f_\theta \in \mathcal{H}} \widehat{L}(f_\theta, \widehat{\lambda}^\star_\Phi). \tag{14}$$

Then, with probability at least $1 - 4\delta$, $\widehat{f}_{\theta_\Phi}$ is near-feasible and near-optimal for problem equation P. In particular,

$$\mathbb{E}_{z\sim D}\left[[\ell(\widehat{f}_{\theta_\Phi}, z)]_+\right] \leq \sqrt{\frac{2L^2}{\mu}\left[\Delta_D + \zeta_0 + L(1+\Gamma)\nu_{\mathcal{H}}\right]} + \sqrt{\frac{2L^3(1+\Gamma)\nu_{\mathcal{H}}}{\mu}} + \zeta =: \Delta_F, \tag{15}$$

$$\left|\mathbb{E}_{z\sim D_0}\left[\ell_0(\widehat{f}_{\theta_\Phi}, z)\right] - P^\star\right| \leq \zeta_0 + \Delta_D + L\left(1 + \frac{B_0}{s}\right)\nu_{\mathcal{H}} + \left(\Gamma + \frac{B\mu}{L^2}\right)\Delta_F =: \Delta_O. \tag{16}$$

Proof. See Appendix A.3. □

Theorem 1 shows that solving the parametrized empirical dual recovers a primal model whose statistical constraint violation and objective error are proportional to either the statistical, dual-parameterization, or model-approximation errors. In particular, whenever these three approximations simultaneously vanish, that is $\zeta_0, \zeta, \nu_\Phi, \nu_{\mathcal{H}} \to 0$, both the feasibility bound Δ_F and the optimality bound Δ_O vanish to zero. Moreover, we remark that the explicit approximation penalty introduced by restricting the functional dual domain to Λ_Φ is only the error $B\nu_\Phi$ from approximating the optimal multiplier. This error propagates to primal feasibility and optimality only through Δ_D. The remaining terms arise from finite-sample estimation and from recovering a primal solution in the original, possibly non-convex, hypothesis class. This implies that as the parametric dual family becomes expressive enough to approximate the optimal statistical multiplier, its primal recovery guarantees approach those obtained from the corresponding unrestricted empirical dual formulation.

3.2 Sensitivity of the learned parametric multiplier

The primal recovery guarantees of Theorem 1 establish that the parametrized empirical dual remains a meaningful surrogate for the statistical primal problem P. Our main motivation for learning a parametric multiplier, however, is to retain the sensitivity interpretation of the optimal statistical multiplier $\lambda^\star$ while being able to evaluate it on new, unseen samples. We formalize this interpretation next.

Let $u : \mathcal{Z} \to \mathbb{R}$ assign a pointwise perturbation to the constraint threshold. We define the perturbation function $P(u)$ as the optimal value of the perturbed primal problem,

$$\begin{aligned} P(u) = \min_{f_\theta \in \mathcal{H}} \quad & \mathbb{E}_{z\sim\mathcal{D}_0}[\ell_0(f_\theta, z)] \\ \text{s. to} \quad & \ell(f_\theta, z) \leq u(z) \qquad \mathcal{D}\text{-a.e.} \end{aligned} \tag{P_u}$$

The perturbation u changes the original constraint threshold from 0 to $u(z)$ at each sample. Since larger values of u relax the constraints, $P(u)$ is nonincreasing with respect to the pointwise order. The optimal statistical multiplier characterizes the sensitivity of this value function.

Lemma 2 (Sensitivity of the optimal multiplier)**.** *Let $\lambda^\star$ be the optimal statistical dual multiplier. Then $-\lambda^\star$ is a Δ_{gap}-subgradient of $P(0) = P^*$.*

That is, for every perturbation u

$$P(u) \geq P^* - \mathbb{E}_{z\sim\mathcal{D}}[\lambda^\star(z)u(z)] - \Delta_{\mathrm{gap}}, \tag{17}$$

where $\Delta_{\mathrm{gap}} := |P^\star - D^\star|$ is the duality gap.

Lemma 2 shows that the multiplier $\lambda^\star(z)$ quantifies how costly it can be to tighten a constraint associated with a particular sample. In particular, a larger multiplier will potentially translate into a larger increase in the optimal objective when that constraint is tightened, whereas a zero multiplier implies no decrease from relaxing it will happen to the objective. As a learned parametric multiplier can be evaluated on unseen samples, it provides a measure of how much enforcing the constraint on an unseen sample—for instance, by including that sample in the training set—can affect the optimal objective.

To transfer this interpretation to the learned multiplier we require also that the parametric dual class generalizes.

Assumption 7 (Dual class complexity)**.** *There exists a function $\zeta_\Lambda(N, \delta)$, monotonically decreasing to zero in N, such that, with probability at least $1 - \delta$,*

$$\sup_{\lambda_\phi\in\Lambda_\Phi} \left| \mathbb{E}_D[\lambda_\phi(z)] - \frac{1}{N}\sum_{z_n\in S} \lambda_\phi(z_n) \right| \leq \zeta_\Lambda(N, \delta). \tag{18}$$

Assumption 7 mirrors Assumption 1 for the dual class. As before, we assume that $\zeta_\Lambda(N, \delta)$ arises from vanishing Rademacher complexity and refer to Appendix A.1 for the details. Under the same assumption requiered primal recover, with the dual complexity bound, we can characterize the sensitivity of the learned multiplier.

Theorem 2 (Empirical approximate sensitivity)**.** *Let assumptions 1- 7 hold. Then, with probability at least $1 - 4\delta$, for every $u \in L_1(D)$ it holds that*

$$P(u) \geq P(0) - \mathbb{E}_{z\sim D}\left[\widehat{\lambda}^\star_\Phi(z)u(z)\right] - \left[L\left(1 + \frac{B_0}{s}\right)\nu_\mathcal{H} + B\nu_\Phi + 2\zeta_0 + 2\Gamma\zeta + B\zeta_\Lambda\right]. \tag{19}$$

Proof. See Appendix A.4. □

Theorem 2 shows that the learned multiplier is also an approximate subgradient of the perturbation function. In this case, however, additional terms appear in the subgradient approximation. Specifically, the term involving $\nu_\mathcal{H}$ upper-bounds the duality gap, while $B\nu_\Phi$ captures the approximation error induced by the parametric dual family. The remaining terms arise from finite-sample estimation, with $B\zeta_\Lambda$ accounting specifically for the statistical complexity of the learned multiplier class.

The result also reveals a tradeoff in the choice of Λ_Φ. A richer dual class may reduce the approximation error ν_Φ, but at the cost of greater statistical complexity and, consequently, a larger ζ_Λ. Sensitivity learning therefore favors a class that is expressive enough to approximate $\lambda^\star$ while remaining statistically learnable. This tradeoff does not arise in the same form in Theorem 1: primal recovery depends on the dual class through ν_Φ, whereas the additional complexity term ζ_Λ appears only when the learned multiplier itself must generalize to unseen samples.

In the next section, we empirically study the phenomena predicted by the theory and study how the expressivity of the parametric dual family affects the sensitivity estimation. Our experiments use an augmented Lagrangian formulation, for which the theoretical results admit analogous extensions. To keep the main development concise, we defer these extensions to Appendix A.5.

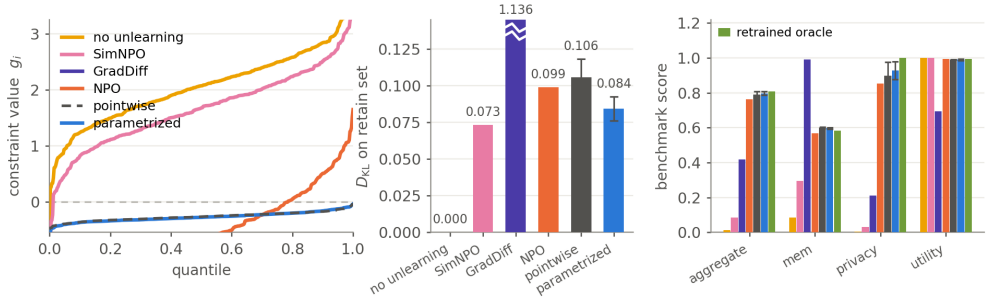


Figure 1: **Primal recovery on TOFU.** *Left:* distribution of forgetting-constraint values, where negative values indicate constraint satisfaction. *Center:* KL divergence on the retain set, measuring deviation from the original model. *Right:* aggregate unlearning performance across memorization, privacy, and utility metrics. *The parameterized dual closely matches the pointwise in constraint satisfaciton, objective value and downstream performance.*

4 Numerical Experiments

In this section, we evaluate the proposed parametric dual formulation across four diverse learning problems. We consider image classification, agentic workflow prediction, language-model unlearning, and optimal power flow. We next provide a brief description of each task and its evaluation protocol; complete implementation details are provided in Appendix D.

MNIST: Image classification. We consider robust classification on MNIST (Lecun et al., 1998), with 60,000 handwritten digit images. Each sample is constrained to be correctly classified with a prescribed margin, while the primal objective minimizes the squared ℓ_2 norm of the model parameters. Both the primal predictor and dual multiplier are parameterized by CNNs.

FLORA-Bench: Agentic workflow prediction. We use the Coding-GD dataset from FLORA-Bench (Zhang et al., 2025), where a graph neural network predicts whether a multi-LM workflow solves a coding task. The primal objective optimizes task-wise ranking, while pointwise binary cross-entropy constraints enforce accurate classification for every workflow–task pair. Both the primal predictor and the dual multiplier are parameterized by lightweight graph neural networks with a shared backbone.

TOFU: Language-model unlearning. We study constrained language-model unlearning on TOFU (Maini et al., 2024) using the OpenUnlearning framework (Dorna et al., 2026). The primal objective preserves model behavior on a retain set, while pointwise constraints enforce forgetting on each example in the forget set. The primal model is a pretrained language model, while the dual multiplier is parameterized by a lightweight neural network over sample representations.

IEEE 30: Optimal power flow prediction. We consider AC optimal power flow prediction on the IEEE 30-bus system following Owerko et al. (2025). The primal model minimizes expected generation cost while enforcing power-flow equations and operational limits for each demand realization. Both the primal predictor and dual multiplier are parameterized by graph neural networks over the power-network topology.

4.1 Primal recovery for the parametric dual

Dual parameterizations preserve primal performance. We first evaluate whether restricting the dual variables to a parametric family degrades the recovered primal solution. Figure 1 reports results on the TOFU task. Both the pointwise and parameterized dual formulations satisfy the forgetting constraints across essentially the entire forget set, unlike the standard unlearning baselines, which exhibit substantial violations. At the same time, the parameterized formulation preserves the retain distribution particularly well, achieving a retain-set KL divergence of 0.084, comparable to the pointwise formulation (0.106) and lower than the remaining unlearning baselines. This behavior

	Optimality Gap (%)		Mean Violation (%)		Max Violation (%)			
	Mean	Std	Mean	Std	Mean	Std	P95	Max
Param	-3.63	1.15	2.70	1.26	6.26	2.61	8.04	15.80
Dual-P	0.18	1.41	0.49	0.62	1.83	2.15	5.65	19.94
Dual-H	-0.08	0.55	1.05	0.30	5.61	0.27	6.03	6.65
Dual-S	2.18	1.25	1.34	0.30	8.22	1.98	11.61	13.47

Table 1: **Primal recovery on the IEEE 30-bus benchmark.** Param replaces the independent per-sample multipliers of Dual-P with a learned dual network conditioned on primal embeddings. Metrics are computed per test operating point and aggregated across the test set. The remaining baselines are reproduced from Owerko et al. (2025), Fig. 10.

is also reflected in the benchmark metrics, where the parameterized and pointwise formulations achieve nearly identical memorization, privacy, and utility scores. These results show that learning the dual variables through a shared parametric function preserves the primal performance of independent pointwise multipliers while enabling the multiplier to generalize to unseen samples. We observe the same behavior in the other dataset in the additional results provided in Appendix D.1.

Primal performance remains stable as the number of constraints grows. We next evaluate the parameterized dual formulation on the IEEE 30-bus benchmark, a standard power-systems test case with 30 buses, 6 generators, and 41 transmission branches, yielding hundreds of constraints per sample. This setting is particularly challenging for a shared dual parameterization, since different constraint types act on distinct parts of the network and their associated multipliers depend on the complete operating point. Nevertheless, Table 1 shows that the resulting primal GNN remains competitive with standard empirical dual formulations. The parameterized model achieves a negative mean optimality gap of -3.63%, while maintaining constraint violations of the same overall scale as the pointwise-dual baselines. These results suggest that a single learned multiplier function can accommodate hundreds of heterogeneous pointwise constraints without substantially degrading the quality of the recovered primal solution.

4.2 Learnability of the dual multipliers

Learned multipliers generalize to unseen samples. We next evaluate whether the learned multiplier preserves dual information on unseen samples. Table 2 compares predicted multipliers with an independently computed pointwise reference solution. Across MNIST, OPF, and FLORA-Bench, the learned multiplier reliably identifies active constraints, with tight-set AUCs between 0.877 and 0.935, and ranks the most important constraints well, with NDCG between 0.836 and 0.914. Global Spearman correlations are lower, as pointwise multipliers are highly sparse and concentrated near zero for inactive constraints, making their relative ordering poorly conditioned. TOFU presents a different regime, with about 80% of constraints active, leaving fewer inactive samples for separation and less variation across multiplier values. Ranking performance therefore degrades, although

Dataset	Active-set base rate	Dual recovery on unseen samples		
		Tight-set AUC ↑	NDCG ↑	Spearman ρ ↑
MNIST	0.47%	0.923 ± 0.013	0.836 ± 0.036	0.277 ± 0.026
OPF	5.13%	0.877 ± 0.063	0.902 ± 0.008	0.324 ± 0.035
TOFU	80.00%	0.713 ± 0.078	0.386 ± 0.040	-0.048 ± 0.051
FLORA-Bench	54.08%	0.935 ± 0.001	0.914 ± 0.001	0.211 ± 0.008

Table 2: **Generalization of learned dual multipliers to unseen samples.** Predicted multipliers are evaluated against an independently computed pointwise dual reference. Tight-set AUC measures identification of active constraints, while NDCG and Spearman's ρ measure recovery of the multiplier ranking. Results are reported as mean $\pm$ standard deviation.

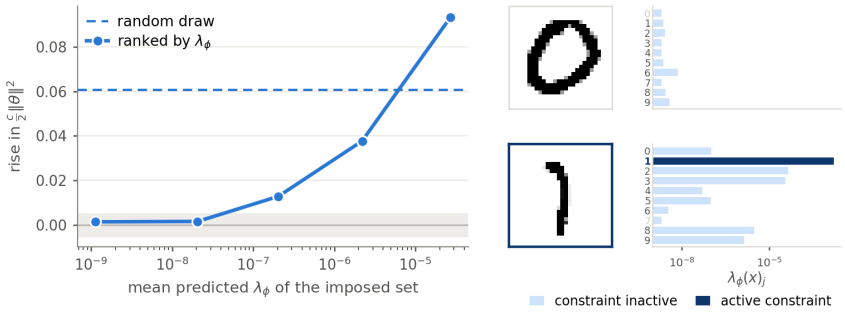


Figure 2: **Sensitivity estimation on MNIST.** *Left:* objective increase after imposing held-out constraints versus their mean predicted multiplier. Larger λ_ϕ values induce larger increases than random draws. *Right:* representative low- and high-sensitivity samples with their classwise multipliers.

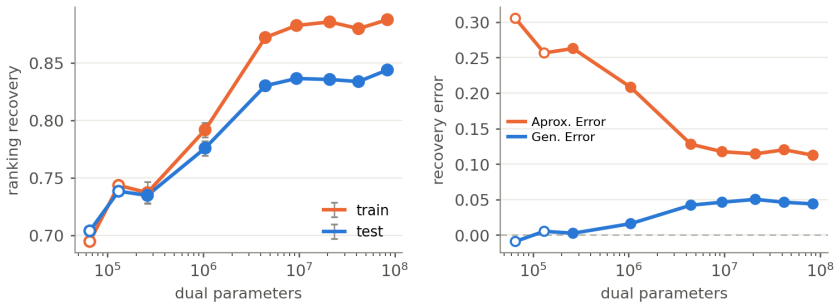


Figure 3: **Effect of dual parameterization capacity on FLORA-Bench.** *Left:* ranking recovery on the training and test sets as the number of dual-network parameters increases. *Right:* decomposition into approximation and generalization errors. Increasing capacity substantially reduces approximation error and improves sensitivity recovery, while introducing a growing generalization error, illustrating the tradeoff between dual expressivity and statistical complexity.

active-set identification remains nontrivial, with AUC 0.713. Overall, the parametric dual transfers meaningful constraint-importance information to unseen samples, particularly through active-set identification and top-ranked sensitivity.

Multiplier magnitude captures sample-level sensitivity. We test whether the learned multiplier predicts how strongly an unseen sample constrains the optimization problem. On MNIST, we form held-out sets with different average predicted multipliers and re-solve the problem after imposing their pointwise constraints. Figure 2 shows that the objective increase grows monotonically with the mean predicted multiplier, while random sets produce a nearly constant baseline. This agrees with the sensitivity interpretation in Section 3.2: larger λ_ϕ identifies constraints whose enforcement has a greater effect on the optimum. The examples on the right illustrate this behavior: an unambiguous zero receives uniformly small multipliers, whereas an ambiguous seven resembling a one receives a large multiplier for the corresponding class constraint. Such samples require greater classification capacity, explaining their larger effect on the objective. Additional FLORA-Bench results are provided in Appendix D.3.

Dual capacity balances approximation and generalization. We study how dual-model capacity affects sensitivity recovery. Figure 3 shows that increasing the number of dual parameters improves ranking recovery on both training and test sets, indicating reduced approximation error, but the gains saturate beyond roughly 10^7 parameters. The right panel makes the tradeoff explicit: approximation error decreases sharply, while generalization error initially grows and then levels off. This empirically matches Theorem 2: richer dual classes better approximate the optimal multiplier but incur greater statistical complexity. Additional MNIST capacity results are provided in Appendix D.2.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation of experimental methods, the development and editing of software code, the selection and refinement of experimental parameters, and aspects of experimental methodology. We also used generative AI tools to assist in the verification of mathematical arguments and proofs. AI-assisted mathematical content was independently reviewed and verified by the authors.

We additionally used generative AI tools to assist with improving the clarity, grammar and readability organization of the manuscript, and identifying relevant literature and references. All AI-assisted code was manually reviewed and tested for correctness, and all AI-assisted mathematical claims, proofs, experimental conclusions, and citations were checked by the authors.

We did not use generative AI tools to generate synthetic datasets, formulate the central theoretical models or conceptual framework of the work, formulate the main mathematical claims, propose or refine the primary research hypotheses, translate research materials, clean or reformat datasets, or perform qualitative or thematic data analysis.

The authors have reviewed all AI-assisted work and take responsibility for the final content of this work, including all text, mathematical claims, proofs, code, experimental results, citations, and other artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Given the nature of the manuscript, we do not believe an additional ethic statement is needed.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide one ready-to-run repository for each task, enabling reproduction of the experiments and results reported throughout the manuscript. The repositories are designed to be as simple and self-contained as possible and include comments to improve clarity and ease of use.

Regarding the theoretical results, complete proofs of all theoretical statements are provided in the Appendix, with additional intermediate steps and explanations to facilitate their verification.

REFERENCES

Yoshua Bengio, Nicolas Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. In Y. Weiss, B. Schölkopf, and J. Platt (eds.), *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/0fc170ecbb8ff1afb2c6de48ea5343e7-Paper.pdf.

Daniel Bienstock and Abhinav Verma. Strong NP-hardness of AC power flows feasibility. *Operations Research Letters*, 47(6):494–501, 2019.

Ignacio Boero, Ignacio Hounie, Luiz Chamon, and Alejandro Ribeiro. Everywhere learning: Artificial intelligence with pointwise constraints. *arXiv preprint arXiv:2606.01557*, 2026.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Mary B. Cain, Richard P. O'Neill, and Anya Castillo. History of optimal power flow and formulations. Technical report, Federal Energy Regulatory Commission, 2012. Increasing Efficiency through Improved Software, pp. 1–31.

Anya Castillo and Richard P. O'Neill. Computational performance of solution techniques applied to the ACOPF. Technical report, Federal Energy Regulatory Commission, 2013. Increasing Efficiency through Improved Software, pp. 1–34.

Luiz F. O. Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Luiz F. O. Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning, 2021. URL https://arxiv.org/abs/2006.05487.

Spyros Chatzivasileiadis. Lecture notes on optimal power flow (OPF), 2018.

Kejun Chen, Shourya Bose, and Yu Zhang. Unsupervised deep learning for AC optimal power flow via Lagrangian duality. In *IEEE Global Communications Conference (GLOBECOM)*, pp. 5305–5310, 2022.

Carleton Coffrin, Russell Bent, Kaarthik Sundar, Yeesian Ng, and Miles Lubin. PowerModels.jl: An open-source framework for exploring power flow formulations. In *Power Systems Computation Conference (PSCC)*, pp. 1–8, 2018.

Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. *Journal of Machine Learning Research*, 20(172):1–59, 2019. URL http://jmlr.org/papers/v20/18-616.html.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zico Kolter, Zachary Lipton, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *Advances in Neural Information Processing Systems*, 38, 2026.

Ferdinando Fioretto, Terrence W. K. Mak, and Pascal Van Hentenryck. Predicting AC optimal power flows: Combining deep learning and Lagrangian dual methods. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 630–637, 2020.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks, 2019. URL https://arxiv.org/abs/1712.06541.

Neel Guha, Zhecheng Wang, and Arun Majumdar. Machine learning for AC optimal power flow. In *36th International Conference on Machine Learning (ICML)*, 2019.

Ignacio Hounie, Ignacio Boero, and Alejandro Ribeiro. Every sample counts: Supervised fine-tuning of language models with pointwise constraints. *arXiv preprint arXiv:2607.08968*, 2026.

Wanjun Huang, Xiang Pan, Minghua Chen, and Steven H. Low. DeepOPF-V: Solving AC-OPF problems efficiently. *IEEE Transactions on Power Systems*, 37(1):800–803, 2022.

Wanjun Huang, Minghua Chen, and Steven H. Low. Unsupervised learning for solving AC optimal power flows: Design, analysis, and experiment. *IEEE Transactions on Power Systems*, 39(6): 7102–7114, 2024.

Burak Kocuk, Santanu S. Dey, and X. Andy Sun. Strong SOCP relaxations for the optimal power flow problem. *Operations Research*, 64(6):1177–1196, 2016.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Michel. Ledoux and Michel. Talagrand. *Probability in Banach Spaces : Isoperimetry and Processes*. Ergebnisse der Mathematik und ihrer Grenzgebiete, A Series of Modern Surveys in Mathematics, 23. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991. ISBN 9783642202124.

Bernard Lesieutre and Ian Hiskens. Convexity of the set of feasible injections and revenue adequacy in FTR markets. *IEEE Transactions on Power Systems*, 20(4):1790–1798, 2005.

Steven H. Low. Convex relaxation of optimal power flow—part II: Exactness. *IEEE Transactions on Control of Network Systems*, 1(2):177–189, 2014.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. Tofu: A task of fictitious unlearning for llms, 2024. URL https://arxiv.org/abs/2401.06121.

Daniel K. Molzahn and Ian A. Hiskens. Convex relaxations of optimal power flow problems: An illustrative example. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 63(5):650–660, 2016.

Damian Owerko, Fernando Gama, and Alejandro Ribeiro. Optimal power flow using graph neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5930–5934, 2020.

Damian Owerko, Fernando Gama, and Alejandro Ribeiro. Unsupervised optimal power flow using graph neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6885–6889, 2024.

Damian Owerko, Anna Scaglione, and Alejandro Ribeiro. Learning optimal power flow with pointwise constraints, 2025.

Xiang Pan, Minghua Chen, Tianyu Zhao, and Steven H. Low. DeepOPF: A feasibility-optimized deep neural network approach for AC optimal power flow problems. *IEEE Systems Journal*, 17 (1):673–683, 2023.

Seonho Park and Pascal Van Hentenryck. Self-supervised primal-dual learning for constrained optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4052–4060, 2023.

R. Tyrrell Rockafellar. Lagrange multipliers and optimality. *SIAM Review*, 35(2):183–238, 1993. doi: 10.1137/1035044. URL https://doi.org/10.1137/1035044.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory, Third Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021. doi: 10.1137/1.9781611976595. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611976595.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021.

A. P. Wierzbicki and S. Kurcyusz. Projection on a cone, penalty functionals and duality theory for problems with inequality constraints in Hilbert space. *SIAM Journal on Control and Optimization*, 15(1):25–56, 1977.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 962–970. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/zafar17a.html.

Yuanshuo Zhang, Yuchen Hou, Bohan Tang, Shuo Chen, Muhan Zhang, Xiaowen Dong, and Siheng Chen. Gnns as predictors of agentic workflow performances, 2025. URL https://arxiv.org/abs/2503.11301.

Ray D. Zimmerman, Carlos E. Murillo-Sánchez, and Robert J. Thomas. MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, 26(1):12–19, 2011.

APPENDIX

The appendix is organized as follows:

- Theoretical Results (Appendix A): Proofs of the theorems on the main body, derivation of uniform convergance rates from Radamacher complexities and extension to Augmented Lagrangian.
- Task Details (Appendix B): Description of the different tasks, baselines, datasets and relevant hyperparameters used in our experiments.
- Implementation Details (Appendix C): Algorithm implementation and primal/dual updates, evaluation of primal recover performance and dual learning.
- Additional Results (Appendix D): Additional figures and tables.

A THEORETICAL RESULTS

A.1 DERIVATION OF UNIFORM CONVERGENCE RATES FROM RADEMACHER COMPLEXITY

In this section, we show that the uniform-convergence rates assumed in Assumptions 1 and 7 follow from standard Rademacher-complexity arguments. We further record the consequences for the weighted and positive-part loss classes used in the proofs.

Let

$$\mathcal{L}_0 := \{z \mapsto \ell_0(f_\theta, z) \ : \ f_\theta \in \mathcal{H}\}, \tag{20}$$

$$\mathcal{L} := \{z \mapsto \ell(f_\theta, z) \ : \ f_\theta \in \mathcal{H}\}, \tag{21}$$

denote the loss classes induced by the hypothesis class.

For a function class $\mathcal{G}$ and an i.i.d. sample $S = \{z_n\}_{n=1}^N$, define its Rademacher complexity as

$$\mathfrak{R}_N(\mathcal{G}) := \mathbb{E}_{S,\sigma}\left[\sup_{g\in\mathcal{G}}\left|\frac{1}{N}\sum_{n=1}^N \sigma_n g(z_n)\right|\right], \tag{22}$$

where $\sigma_1, \ldots, \sigma_N$ are independent Rademacher random variables. Standard symmetrization and concentration arguments imply that, for a class satisfying $\|g\|_\infty \le M$, with probability at least $1-\delta$,

$$\sup_{g\in\mathcal{G}}\left|\mathbb{E}[g(z)] - \frac{1}{N}\sum_{n=1}^N g(z_n)\right| \le 2\mathfrak{R}_N(\mathcal{G}) + M\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{23}$$

Consequently, the following rates satisfy Assumption 1:

$$\zeta_0(N_0, \delta) := 4\mathfrak{R}_{N_0}(\mathcal{L}_0) + 2B_0\sqrt{\frac{2\log(2/\delta)}{N_0}}, \tag{24}$$

$$\zeta(N, \delta) := 4\mathfrak{R}_N(\mathcal{L}) + 2B\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{25}$$

In particular, if the corresponding Rademacher complexities vanish with the sample size, as is the case for standard neural networks (Golowich et al., 2019), then $\zeta_0(N_0, \delta), \zeta(N, \delta) \to 0$.

Rademacher-complexity bounds can be transferred directly to several loss classes that arise in our analysis. In particular, Lipschitz transformations and bounded reweightings preserve uniform convergence, up to the corresponding Lipschitz or magnitude factors. We formalize these consequences using the Rademacher contraction inequality.

Lemma 3 (Rademacher contraction inequality)**.** *Let $\mathcal{F}$ be a class of real-valued functions and let $\psi : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz with $\psi(0) = 0$. Then, for any sample $S = (x_1, \ldots, x_N)$,*

$$\widehat{\mathfrak{R}}_S(\psi \circ \mathcal{F}) \le L\,\widehat{\mathfrak{R}}_S(\mathcal{F}).$$

Consequently,

$$\mathfrak{R}_N(\psi \circ \mathcal{F}) \leq L\, \mathfrak{R}_N(\mathcal{F}).$$

Proof. See (Ledoux & Talagrand, 1991, Theorem 4.12). □

The following proposition collects the resulting uniform-convergence bounds used throughout the proof of the theorems.

Proposition 1 (Uniform convergence for derived loss classes)**.** *Let Assumption 1 hold, and let the uniform convergence rate be derived from 25. Then the following hold.*

1. ***Positive-part loss.*** *Let*

$$\mathcal{L}_+ := \{z \mapsto [\ell(f_\theta, z)]_+ : f_\theta \in \mathcal{H}\}.$$

Then

$$\mathfrak{R}_N(\mathcal{L}_+) \leq \mathfrak{R}_N(\mathcal{L}),$$

and, with probability at least $1-\delta$*,*

$$\sup_{f_\theta \in \mathcal{H}} \left| \mathbb{E}_{\mathcal{D}}\big[[\ell(f_\theta, z)]_+\big] - \frac{1}{N} \sum_{z_n \in S} [\ell(f_\theta, z_n)]_+ \right| \leq \zeta(N, \delta). \tag{26}$$

2. ***Fixed weighted loss.*** *Let* $\lambda : \mathcal{Z} \to \mathbb{R}_+$ *be fixed independently of the sample and satisfy* $\|\lambda\|_\infty \leq \Gamma$*. Define*

$$\lambda\mathcal{L} := \{z \mapsto \lambda(z)\ell(f_\theta, z) : f_\theta \in \mathcal{H}\}.$$

Then

$$\mathfrak{R}_N(\lambda\mathcal{L}) \leq \Gamma\, \mathfrak{R}_N(\mathcal{L}),$$

and, with probability at least $1-\delta$*,*

$$\sup_{f_\theta \in \mathcal{H}} \left| \mathbb{E}_{\mathcal{D}}[\lambda(z)\ell(f_\theta, z)] - \frac{1}{N} \sum_{z_n \in S} \lambda(z_n)\ell(f_\theta, z_n) \right| \leq \Gamma\, \zeta(N, \delta). \tag{27}$$

3. ***Learned dual class.*** *Suppose* $\|\lambda_\phi\|_\infty \leq \Gamma$ *for every* $\lambda_\phi \in \Lambda_\Phi$ *and let Assumption 7 hold with a rate of the form*

$$\zeta_\Lambda(N, \delta) := 4\mathfrak{R}_N(\Lambda_\Phi) + 2\Gamma\sqrt{\frac{2\log(2/\delta)}{N}}. \tag{28}$$

Then, for the product class

$$\Lambda_\Phi\mathcal{L} := \{z \mapsto \lambda_\phi(z)\ell(f_\theta, z) : \lambda_\phi \in \Lambda_\Phi,\ f_\theta \in \mathcal{H}\},$$

it holds that

$$\mathfrak{R}_N(\Lambda_\Phi\mathcal{L}) \leq 2\left[\Gamma\mathfrak{R}_N(\mathcal{L}) + B\mathfrak{R}_N(\Lambda_\Phi)\right],$$

and thus, with probability at least $1-\delta$*,*

$$\sup_{\substack{\lambda_\phi \in \Lambda_\Phi \\ f_\theta \in \mathcal{H}}} \left| \mathbb{E}_{\mathcal{D}}[\lambda_\phi(z)\ell(f_\theta, z)] - \frac{1}{N} \sum_{z_n \in S} \lambda_\phi(z_n)\ell(f_\theta, z_n) \right| \leq \Gamma\zeta(N, \delta) + B\zeta_\Lambda(N, \delta). \tag{29}$$

Proof. For the positive-part loss, the map $t \mapsto [t]_+$ is 1-Lipschitz, so the Rademacher contraction inequality in 3 gives

$$\mathfrak{R}_N(\mathcal{L}_+) \leq \mathfrak{R}_N(\mathcal{L}).$$

Since $[\ell(f_\theta, z)]_+ \leq B$, applying the uniform-convergence bound in equation 23 yields equation 26.

For the fixed weighted loss, $t \mapsto \lambda(z_n)t$ is Γ-Lipschitz. Applying the contraction inequality in 3 therefore gives

$$\mathfrak{R}_N(\lambda\mathcal{L}) \leq \Gamma\mathfrak{R}_N(\mathcal{L}).$$

Together with $|\lambda(z)\ell(f_\theta, z)| \leq \Gamma B$, equation 23 gives equation 27.

Finally, the jointly learned multiplier and loss, the standard Rademacher product-class inequality gives

$$\mathfrak{R}_N(\Lambda_\Phi\mathcal{L}) \leq 2\left[\Gamma\mathfrak{R}_N(\mathcal{L}) + B\mathfrak{R}_N(\Lambda_\Phi)\right].$$

Since $|\lambda_\phi(z)\ell(f_\theta, z)| \leq \Gamma B$, applying equation 23 and substituting the definitions of $\zeta(N, \delta)$ and $\zeta_\Lambda(N, \delta)$ yields equation 29. □

A.2 Proof of Lemma 1

Proof. Let the empirical dual function be

$$\widehat{g}(\lambda) := \inf_{f_\theta \in \mathcal{H}} \left\{ \frac{1}{N_0} \sum_{n=1}^{N_0} \ell_0(f_\theta, z_n) + \frac{1}{N} \sum_{n=1}^{N} \lambda(z_n) \ell(f_\theta, z_n) \right\}. \tag{30}$$

We first bound $D^\star - \widehat{D}^\star_\Phi$. For any $\epsilon > 0$, by Assumption 2, there exists a fixed $\bar{\lambda}_\phi \in \Lambda_\Phi$ such that

$$\|\bar{\lambda}_\phi - \lambda^\star\|_{L_1(D)} \le \nu_\Phi + \epsilon. \tag{31}$$

Since $|\ell(f_\theta, z)| \le B$, for every $f_\theta \in \mathcal{H}$,

$$\left| \mathbb{E}_D[\bar{\lambda}_\phi(z)\ell(f_\theta, z)] - \mathbb{E}_D[\lambda^\star(z)\ell(f_\theta, z)] \right| \le B(\nu_\Phi + \epsilon). \tag{32}$$

Taking the infimum over f_θ gives

$$g(\bar{\lambda}_\phi) \ge D^\star - B(\nu_\Phi + \epsilon). \tag{33}$$

Because $\bar{\lambda}_\phi$ is fixed independently of the sample and satisfies $\|\bar{\lambda}_\phi\|_\infty \le \Gamma$, the weighted uniform-convergence result of Proposition 1 yields

$$\widehat{g}(\bar{\lambda}_\phi) \ge g(\bar{\lambda}_\phi) - \zeta_0(N_0, \delta) - \Gamma\zeta(N, \delta). \tag{34}$$

Since $\widehat{D}^\star_\Phi$ maximizes the empirical dual over Λ_Φ,

$$\begin{aligned} \widehat{D}^\star_\Phi &\ge \widehat{g}(\bar{\lambda}_\phi) && (35)\\ &\ge D^\star - B(\nu_\Phi + \epsilon) - \zeta_0(N_0, \delta) - \Gamma\zeta(N, \delta). && (36) \end{aligned}$$

Letting $\epsilon \to 0$ gives

$$D^\star - \widehat{D}^\star_\Phi \le B\nu_\Phi + \zeta_0(N_0, \delta) + \Gamma\zeta(N, \delta). \tag{37}$$

For the opposite direction, let $\widehat{D}^\star$ denote the unrestricted empirical dual optimum obtained by assigning an independent nonnegative multiplier to each constraint sample. Since the evaluations of every $\lambda_\phi \in \Lambda_\Phi$ form a feasible multiplier vector for this unrestricted problem,

$$\widehat{D}^\star_\Phi \le \widehat{D}^\star. \tag{38}$$

We next relate the unrestricted empirical dual to the statistical dual. Fix an empirical multiplier vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^N_+$ and $\varepsilon > 0$. Under mild regularity of the constraint sample space,[2] one can construct $\lambda_\varepsilon \in L^+_1(D)$ such that

$$\sup_{f_\theta \in \mathcal{H}} \left| \mathbb{E}_D[\lambda_\varepsilon(z)\ell(f_\theta, z)] - \frac{1}{N} \sum_{n=1}^{N} \lambda_n \ell(f_\theta, z_n) \right| \le \varepsilon. \tag{39}$$

Indeed, this follows by placing the mass λ_n/N on sufficiently small neighborhoods of each z_n.

Let $\widehat{g}(\boldsymbol{\lambda})$ denote the empirical dual value associated with $\boldsymbol{\lambda}$. On the uniform-convergence event for the objective,

$$\begin{aligned} \widehat{g}(\boldsymbol{\lambda}) &= \inf_{f_\theta \in \mathcal{H}} \left\{ \frac{1}{N_0} \sum_{z_n \in S_0} \ell_0(f_\theta, z_n) + \frac{1}{N} \sum_{n=1}^{N} \lambda_n \ell(f_\theta, z_n) \right\} && (40)\\ &\le \inf_{f_\theta \in \mathcal{H}} \{ \mathbb{E}_{D_0}[\ell_0(f_\theta, z)] + \mathbb{E}_D[\lambda_\varepsilon(z)\ell(f_\theta, z)] \} + \zeta_0(N_0, \delta) + \varepsilon && (41)\\ &= g(\lambda_\varepsilon) + \zeta_0(N_0, \delta) + \varepsilon && (42)\\ &\le D^\star + \zeta_0(N_0, \delta) + \varepsilon. && (43) \end{aligned}$$

[2] Specifically, it suffices that the constraint losses vary uniformly continuously with the sample and that D assigns positive mass to sufficiently small neighborhoods of points in its support. For example, this holds when $\mathcal{Z}$ is a compact metric space and $z \mapsto \ell(f_\theta, z)$ is uniformly Lipschitz over $f_\theta \in \mathcal{H}$. We omit these topological regularity conditions from the main statement to keep the presentation focused on the statistical and approximation aspects of the result.

Since this holds for every empirical multiplier vector and every $\varepsilon > 0$, taking the supremum over $\boldsymbol{\lambda} \in \mathbb{R}^N_+$ and then letting $\varepsilon \to 0$ yields

$$\widehat{D}^\star \le D^\star + \zeta_0(N_0, \delta). \tag{44}$$

Consequently,

$$\widehat{D}^\star_\Phi - D^\star \le \zeta_0(N_0, \delta). \tag{45}$$

Combining equation 37 and equation 45, and applying a union bound over the corresponding uniform-convergence events, yields, with probability at least $1 - 3\delta$,

$$\left| D^\star - \widehat{D}^\star_\Phi \right| \le B\nu_\Phi + \zeta_0(N_0, \delta) + \Gamma\zeta(N, \delta) = \Delta_D. \tag{46}$$

□

A.3 Proof of Theorem 1

Proof. We start by defining notation.

Lets define the convexified empirical primal problem as

$$\widehat{P}^\star_{\mathcal{C}} := \min_{f \in \overline{\mathcal{H}}} \frac{1}{N_0} \sum_{z_n \in S_0} \ell_0(f, z_n) \qquad \text{s.t.} \qquad \ell(f, z_n) \le 0, \quad z_n \in S. \tag{47}$$

For $u, v \in \mathbb{R}^N$, write

$$\langle u, v \rangle_N := \frac{1}{N} \sum_{n=1}^N u_n v_n, \qquad \|u\|^2_{2,N} := \frac{1}{N} \sum_{n=1}^N u_n^2, \tag{48}$$

and let

$$\ell_S(f) := \big(\ell(f, z_1), \ldots, \ell(f, z_N)\big). \tag{49}$$

The convexified empirical dual function is

$$\widehat{g}_{\mathcal{C}}(\lambda) := \min_{f \in \overline{\mathcal{H}}} \left\{ \frac{1}{N_0} \sum_{z_n \in S_0} \ell_0(f, z_n) + \langle \lambda, \ell_S(f) \rangle_N \right\}, \qquad \lambda \in \mathbb{R}^N_+. \tag{50}$$

Let

$$\widehat{D}^\star_{\mathcal{C}} := \max_{\lambda \ge 0} \widehat{g}_{\mathcal{C}}(\lambda), \tag{51}$$

and denote by

$$\widehat{f}_{\mathcal{C}}(\lambda) := \arg\min_{f \in \overline{\mathcal{H}}} \left\{ \frac{1}{N_0} \sum_{z_n \in S_0} \ell_0(f, z_n) + \langle \lambda, \ell_S(f) \rangle_N \right\} \tag{52}$$

the corresponding Lagrangian minimizer. By strong convexity of the objective, this minimizer is unique. The strict Slater condition implies strong duality for the convexified empirical problem, so if $\lambda^\star_{\mathcal{C}}$ is an optimal multiplier, then

$$\widehat{f}^\star_{\mathcal{C}} := \widehat{f}_{\mathcal{C}}(\lambda^\star_{\mathcal{C}}) \tag{53}$$

is feasible for the convexified empirical primal problem.

Finally, let

$$\widehat{\lambda}^\star_\Phi := \big(\widehat{\lambda}^\star_\Phi(z_1), \ldots, \widehat{\lambda}^\star_\Phi(z_N)\big) \tag{54}$$

denote the vector obtained by evaluating the learned parametric multiplier on the constraint sample.

Feasibility. We decompose the discrepancy between the constraint losses of the returned predictor and those of the optimal convexified solution as

$$\left\|\ell_S(\widehat{f}_{\theta_\Phi}) - \ell_S(\widehat{f}^\star_{\mathcal{C}})\right\|_{2,N} \le \left\|\ell_S(\widehat{f}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi)) - \ell_S(\widehat{f}^\star_{\mathcal{C}})\right\|_{2,N} + \left\|\ell_S(\widehat{f}_{\theta_\Phi}) - \ell_S(\widehat{f}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi))\right\|_{2,N}. \tag{55}$$

1) Dual suboptimality and smoothness. Strong convexity of the empirical objective and Lipschitz continuity of the constraint losses imply that $\widehat{g}_{\mathcal{C}}$ is differentiable and

$$\nabla \widehat{g}_{\mathcal{C}}(\lambda) = \ell_S(\widehat{f}_{\mathcal{C}}(\lambda)). \tag{56}$$

Moreover, its gradient is L^2/μ-Lipschitz. Indeed, for any $\lambda, \lambda' \ge 0$, strong convexity of the corresponding Lagrangians and Lipschitz continuity of the constraint map yield

$$\mu \left\|\widehat{f}_{\mathcal{C}}(\lambda) - \widehat{f}_{\mathcal{C}}(\lambda')\right\|^2 \le \left\langle \lambda - \lambda', \ell_S(\widehat{f}_{\mathcal{C}}(\lambda')) - \ell_S(\widehat{f}_{\mathcal{C}}(\lambda))\right\rangle_N, \tag{57}$$

and therefore

$$\|\nabla \widehat{g}_{\mathcal{C}}(\lambda) - \nabla \widehat{g}_{\mathcal{C}}(\lambda')\|_{2,N} \le \frac{L^2}{\mu}\|\lambda - \lambda'\|_{2,N}. \tag{58}$$

We next bound the dual suboptimality of $\widehat{\lambda}^\star_\Phi$ for the convexified empirical dual. Since the convexified primal domain contains $\mathcal{H}$,

$$\widehat{g}_{\mathcal{C}}(\lambda) \le \widehat{g}(\lambda) \tag{59}$$

for every $\lambda \ge 0$. Furthermore, the unrestricted empirical dual dominates the convexified dual, and the lifting argument used in Lemma 1 gives

$$\widehat{D}^\star_{\mathcal{C}} \le \widehat{D}^\star \le D^\star + \zeta_0. \tag{60}$$

Lemma 1 gives

$$\widehat{D}^\star_\Phi \ge D^\star - \Delta_D. \tag{61}$$

Hence,

$$\widehat{D}^\star_{\mathcal{C}} - \widehat{D}^\star_\Phi \le \Delta_D + \zeta_0. \tag{62}$$

It remains to account for the difference between minimizing the Lagrangian over $\mathcal{H}$ and over $\overline{\mathcal{H}}$. Let $f \in \overline{\mathcal{H}}$. By Assumption 5, there exists $f_\theta \in \mathcal{H}$ satisfying $\|f_\theta - f\|_\infty \le \nu_{\mathcal{H}}$. Lipschitz continuity then gives

$$\left|\frac{1}{N_0}\sum_{z_n \in S_0} \ell_0(f_\theta, z_n) - \frac{1}{N_0}\sum_{z_n \in S_0} \ell_0(f, z_n)\right| \le L\nu_{\mathcal{H}}, \tag{63}$$

and

$$|\ell(f_\theta, z_n) - \ell(f, z_n)| \le L\nu_{\mathcal{H}} \tag{64}$$

for every $z_n \in S$. Since $0 \le \widehat{\lambda}^\star_\Phi(z_n) \le \Gamma$,

$$0 \le \widehat{g}(\widehat{\lambda}^\star_\Phi) - \widehat{g}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi) \le L(1+\Gamma)\nu_{\mathcal{H}}. \tag{65}$$

Combining equation 62 and equation 65,

$$\widehat{D}^\star_{\mathcal{C}} - \widehat{g}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi) \le \Delta_D + \zeta_0 + L(1+\Gamma)\nu_{\mathcal{H}}. \tag{66}$$

Since $-\widehat{g}_{\mathcal{C}}$ is convex and L^2/μ-smooth, cocoercivity gives

$$\left\|\nabla \widehat{g}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi) - \nabla \widehat{g}_{\mathcal{C}}(\lambda^\star_{\mathcal{C}})\right\|^2_{2,N} \le \frac{2L^2}{\mu}\left[\widehat{D}^\star_{\mathcal{C}} - \widehat{g}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi)\right]. \tag{67}$$

Using $\nabla \widehat{g}_{\mathcal{C}}(\lambda) = \ell_S(\widehat{f}_{\mathcal{C}}(\lambda))$ and equation 66, we obtain

$$\left\|\ell_S(\widehat{f}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi)) - \ell_S(\widehat{f}^\star_{\mathcal{C}})\right\|_{2,N} \leq \sqrt{\frac{2L^2}{\mu}\left[\Delta_D + \zeta_0 + L(1+\Gamma)\nu_{\mathcal{H}}\right]}. \tag{68}$$

2) Transfer to the original hypothesis class. By equation 65 and the fact that $\widehat{f}_{\theta_\Phi}$ exactly minimizes the empirical Lagrangian over $\mathcal{H}$,

$$\widehat{L}(\widehat{f}_{\theta_\Phi}, \widehat{\lambda}^\star_\Phi) - \widehat{g}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi) = \widehat{g}(\widehat{\lambda}^\star_\Phi) - \widehat{g}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi) \leq L(1+\Gamma)\nu_{\mathcal{H}}. \tag{69}$$

The convexified Lagrangian is μ-strongly convex. Since $\widehat{f}_{\theta_\Phi} \in \mathcal{H} \subseteq \overline{\mathcal{H}}$,

$$\frac{\mu}{2}\left\|\widehat{f}_{\theta_\Phi} - \widehat{f}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi)\right\|^2 \leq L(1+\Gamma)\nu_{\mathcal{H}}. \tag{70}$$

Lipschitz continuity of the constraint loss therefore gives

$$\left\|\ell_S(\widehat{f}_{\theta_\Phi}) - \ell_S(\widehat{f}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi))\right\|_{2,N} \leq \sqrt{\frac{2L^3(1+\Gamma)\nu_{\mathcal{H}}}{\mu}}. \tag{71}$$

Substituting equation 68 and equation 71 into equation 55 gives

$$\left\|\ell_S(\widehat{f}_{\theta_\Phi}) - \ell_S(\widehat{f}^\star_{\mathcal{C}})\right\|_{2,N} \leq \sqrt{\frac{2L^2}{\mu}\left[\Delta_D + \zeta_0 + L(1+\Gamma)\nu_{\mathcal{H}}\right]} + \sqrt{\frac{2L^3(1+\Gamma)\nu_{\mathcal{H}}}{\mu}}. \tag{72}$$

Since $\widehat{f}^\star_{\mathcal{C}}$ is feasible,

$$\ell(\widehat{f}^\star_{\mathcal{C}}, z_n) \leq 0 \qquad \text{for every } z_n \in S. \tag{73}$$

Therefore,

$$[\ell(\widehat{f}_{\theta_\Phi}, z_n)]_+ \leq \left|\ell(\widehat{f}_{\theta_\Phi}, z_n) - \ell(\widehat{f}^\star_{\mathcal{C}}, z_n)\right|, \tag{74}$$

and hence, by Cauchy–Schwarz,

$$\frac{1}{N}\sum_{z_n \in S}[\ell(\widehat{f}_{\theta_\Phi}, z_n)]_+ \leq \sqrt{\frac{2L^2}{\mu}\left[\Delta_D + \zeta_0 + L(1+\Gamma)\nu_{\mathcal{H}}\right]} + \sqrt{\frac{2L^3(1+\Gamma)\nu_{\mathcal{H}}}{\mu}}. \tag{75}$$

Finally, since $t \mapsto [t]_+$ is 1-Lipschitz and vanishes at the origin, the contraction result of Proposition 1 implies that the positive-part loss has uniform-convergence rate no larger than ζ. Thus,

$$\mathbb{E}_D\left[[\ell(\widehat{f}_{\theta_\Phi}, z)]_+\right] \leq \sqrt{\frac{2L^2}{\mu}\left[\Delta_D + \zeta_0 + L(1+\Gamma)\nu_{\mathcal{H}}\right]} + \sqrt{\frac{2L^3(1+\Gamma)\nu_{\mathcal{H}}}{\mu}} + \zeta = \Delta_F, \tag{76}$$

which proves equation 15.

Optimality. We first control the statistical primal–dual gap. Let $P^\star_{\mathcal{C}}$ denote the optimal value of the statistical primal problem over $\overline{\mathcal{H}}$, and let $P^\star_{\mathcal{C},-L\nu_{\mathcal{H}}}$ denote the same problem with the constraint tightened to

$$\ell(f,z) \le -L\nu_{\mathcal{H}} \qquad D\text{-a.e.} \tag{77}$$

If f is feasible for the tightened convexified problem, Assumption 5 provides $f_\theta \in \mathcal{H}$ satisfying $\|f_\theta - f\|_\infty \le \nu_{\mathcal{H}}$. By Lipschitz continuity, f_θ is feasible for the original statistical problem and its objective differs by at most $L\nu_{\mathcal{H}}$. Hence,

$$P^\star \le P^\star_{\mathcal{C},-L\nu_{\mathcal{H}}} + L\nu_{\mathcal{H}}. \tag{78}$$

By Assumption 6, $\bar{f}_\theta$ satisfies the tightened convexified constraint with margin s. If $\lambda^\star_{\mathcal{C},-L\nu_{\mathcal{H}}}$ is an optimal multiplier of the tightened convexified problem, strong duality and evaluation of its Lagrangian at $\bar{f}_\theta$ give

$$\|\lambda^\star_{\mathcal{C},-L\nu_{\mathcal{H}}}\|_{L_1(D)} \le \frac{B_0}{s}. \tag{79}$$

The sensitivity of the convex value function with respect to tightening the constraint therefore yields

$$P^\star_{\mathcal{C},-L\nu_{\mathcal{H}}} - P^\star_{\mathcal{C}} \le \frac{B_0 L}{s}\nu_{\mathcal{H}}. \tag{80}$$

Since the statistical dual over $\mathcal{H}$ dominates the dual of the convexified problem,

$$D^\star \ge P^\star_{\mathcal{C}}, \tag{81}$$

while weak duality gives $D^\star \le P^\star$. Combining these facts with equation 78,

$$0 \le P^\star - D^\star \le L\left(1 + \frac{B_0}{s}\right)\nu_{\mathcal{H}}. \tag{82}$$

We next control the empirical complementary-slackness term associated with $\widehat{\lambda}^\star_\Phi$. Smoothness of $\widehat{g}_{\mathcal{C}}$ implies that its projected gradient mapping satisfies

$$\left\| \frac{L^2}{\mu}\left[\Pi_{\mathbb{R}^N_+}\left(\widehat{\lambda}^\star_\Phi + \frac{\mu}{L^2}\nabla\widehat{g}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi)\right) - \widehat{\lambda}^\star_\Phi\right]\right\|_{2,N} \le \sqrt{\frac{2L^2}{\mu}\left[\Delta_D + \zeta_0 + L(1+\Gamma)\nu_{\mathcal{H}}\right]}. \tag{83}$$

Using $\nabla\widehat{g}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi) = \ell_S(\widehat{f}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi))$, $0 \le \widehat{\lambda}^\star_\Phi(z_n) \le \Gamma$, and $|\ell| \le B$, the coordinatewise form of the projection gives

$$\begin{aligned}
&\left\|\widehat{\lambda}^\star_\Phi \odot \ell_S(\widehat{f}_{\mathcal{C}}(\widehat{\lambda}^\star_\Phi))\right\|_{1,N} \\
&\qquad \le \max\left\{\Gamma, \frac{B\mu}{L^2}\right\}\sqrt{\frac{2L^2}{\mu}\left[\Delta_D + \zeta_0 + L(1+\Gamma)\nu_{\mathcal{H}}\right]}.
\end{aligned} \tag{84}$$

Transferring from the convexified Lagrangian minimizer to $\widehat{f}_{\theta_\Phi}$ using equation 71 gives

$$\begin{aligned}
&\left\|\widehat{\lambda}^\star_\Phi \odot \ell_S(\widehat{f}_{\theta_\Phi})\right\|_{1,N} \\
&\qquad \le \max\left\{\Gamma, \frac{B\mu}{L^2}\right\}\sqrt{\frac{2L^2}{\mu}\left[\Delta_D + \zeta_0 + L(1+\Gamma)\nu_{\mathcal{H}}\right]} \\
&\qquad\quad + \Gamma\sqrt{\frac{2L^3(1+\Gamma)\nu_{\mathcal{H}}}{\mu}} \\
&\qquad \le \left(\Gamma + \frac{B\mu}{L^2}\right)\Delta_F.
\end{aligned} \tag{85}$$

In particular,

$$\left|\left\langle \widehat{\lambda}^\star_\Phi, \ell_S(\widehat{f}_{\theta_\Phi})\right\rangle_N\right| \le \left(\Gamma + \frac{B\mu}{L^2}\right)\Delta_F. \tag{86}$$

Finally, exact empirical Lagrangian minimization gives

$$\widehat{D}_\Phi^\star = \frac{1}{N_0} \sum_{z_n \in S_0} \ell_0(\widehat{f}_{\theta_\Phi}, z_n) + \left\langle \widehat{\lambda}_\Phi^\star, \ell_S(\widehat{f}_{\theta_\Phi}) \right\rangle_N . \tag{87}$$

Therefore,

$$\begin{aligned} &\mathbb{E}_{D_0}\left[\ell_0(\widehat{f}_{\theta_\Phi}, z)\right] - P^\star \\ &= \left(\mathbb{E}_{D_0}[\ell_0(\widehat{f}_{\theta_\Phi}, z)] - \frac{1}{N_0} \sum_{z_n \in S_0} \ell_0(\widehat{f}_{\theta_\Phi}, z_n)\right) \\ &\quad + \left(\widehat{D}_\Phi^\star - D^\star\right) + (D^\star - P^\star) - \left\langle \widehat{\lambda}_\Phi^\star, \ell_S(\widehat{f}_{\theta_\Phi}) \right\rangle_N . \end{aligned} \tag{88}$$

Applying Assumption 1, Lemma 1, equation 82, and equation 86 yields

$$\begin{aligned} \left|\mathbb{E}_{D_0}\left[\ell_0(\widehat{f}_{\theta_\Phi}, z)\right] - P^\star\right| &\le \zeta_0 + \Delta_D + L\left(1 + \frac{B_0}{s}\right)\nu_{\mathcal{H}} \\ &\quad + \left(\Gamma + \frac{B\mu}{L^2}\right)\Delta_F = \Delta_O, \end{aligned} \tag{89}$$

which proves equation 16.

Union Bound. By a union bound, these events hold simultaneously with probability at least $1-4\delta$.

□

A.4 Proof of Theorem 2

Proof. For the perturbed problem equation P_u, the Lagrangian is

$$L_u(f_\theta, \lambda) = L(f_\theta, \lambda) - \mathbb{E}_D[\lambda(z)u(z)], \tag{90}$$

and therefore its dual function satisfies

$$g_u(\lambda) = g(\lambda) - \mathbb{E}_D[\lambda(z)u(z)]. \tag{91}$$

By weak duality, for every $\lambda \ge 0$,

$$P(u) \ge g(\lambda) - \mathbb{E}_D[\lambda(z)u(z)]. \tag{92}$$

Evaluating at the learned multiplier $\hat{\lambda}_\Phi^\star$ gives

$$P(u) \ge P(0) - \mathbb{E}_D[\hat{\lambda}_\Phi^\star(z)u(z)] - \left[P(0) - g(\hat{\lambda}_\Phi^\star)\right]. \tag{93}$$

The remaining term can be decomposed as

$$\begin{aligned} P(0) - g(\hat{\lambda}_\Phi^\star) &= [P(0) - D^\star] + \left[D^\star - \widehat{D}_\Phi^\star\right] \\ &\quad + \left[\widehat{g}(\hat{\lambda}_\Phi^\star) - g(\hat{\lambda}_\Phi^\star)\right], \end{aligned} \tag{94}$$

where we used

$$\widehat{g}(\hat{\lambda}_\Phi^\star) = \widehat{D}_\Phi^\star. \tag{95}$$

From the primal–dual approximation result used in Theorem 1,

$$P(0) - D^\star \le L\left(1 + \frac{B_0}{s}\right)\nu_{\mathcal{H}}. \tag{96}$$

Lemma 1 gives

$$D^\star - \widehat{D}_\Phi^\star \le B\nu_\Phi + \zeta_0 + \Gamma\zeta. \tag{97}$$

Finally, Proposition 1 gives Λ_Φ and $\mathcal{H}$ gives

$$\left|\widehat{g}(\hat{\lambda}_\Phi^\star) - g(\hat{\lambda}_\Phi^\star)\right| \le \zeta_0 + \Gamma\zeta + B\zeta_\Lambda. \tag{98}$$

Substituting these bounds into equation 94 and then into equation 93 yields equation 19.

□

A.5 Extension to the Augmented Lagrangian

The analysis in Section 3 was developed for the standard Lagrangian. In this section, we show that the same arguments extend to the Powell–Hestenes–Bertsekas (PHB) augmented Lagrangian. The main difference is that augmentation directly provides smoothness of the dual function. Consequently, primal recovery is controlled by the augmentation parameter α, rather than exclusively by the curvature of the primal objective.

For $\alpha > 0$, define the PHB penalty

$$\psi_\alpha(t, \lambda) := \frac{[\lambda + \alpha t]_+^2 - \lambda^2}{2\alpha}, \qquad \lambda \geq 0, \tag{99}$$

and the corresponding statistical augmented Lagrangian

$$\mathcal{L}_\alpha(f_\theta, \lambda) := \mathbb{E}_{\mathcal{D}_0}[\ell_0(f_\theta, z)] + \mathbb{E}_{\mathcal{D}}\left[\psi_\alpha\big(\ell(f_\theta, z), \lambda(z)\big)\right]. \tag{100}$$

Its augmented dual function and optimal value are

$$g_\alpha(\lambda) := \min_{f_\theta \in \mathcal{H}} \mathcal{L}_\alpha(f_\theta, \lambda), \qquad D_\alpha^\star := \sup_{\lambda \in L_1^+(\mathcal{D})} g_\alpha(\lambda). \tag{101}$$

Likewise, from samples $\mathcal{S}_0$ and $\mathcal{S}$, we define

$$\widehat{\mathcal{L}}_\alpha(f_\theta, \lambda) := \frac{1}{N_0} \sum_{z_n \in \mathcal{S}_0} \ell_0(f_\theta, z_n) + \frac{1}{N} \sum_{z_n \in \mathcal{S}} \psi_\alpha\big(\ell(f_\theta, z_n), \lambda(z_n)\big), \tag{102}$$

with

$$\widehat{g}_\alpha(\lambda) := \min_{f_\theta \in \mathcal{H}} \widehat{\mathcal{L}}_\alpha(f_\theta, \lambda). \tag{103}$$

Restricting again the multiplier to Λ_Φ, the parametrized empirical augmented dual problem is

$$\widehat{\lambda}_{\Phi,\alpha}^\star \in \arg\max_{\lambda_\phi \in \Lambda_\Phi} \widehat{g}_\alpha(\lambda_\phi), \qquad \widehat{D}_{\Phi,\alpha}^\star := \widehat{g}_\alpha(\widehat{\lambda}_{\Phi,\alpha}^\star). \tag{104}$$

We use the same assumptions as in the main text. For compactness, let $\zeta_\alpha(N, \delta)$ denote the uniform-convergence error associated with the augmented constraint term in equation 102. Since $\psi_\alpha(t, \lambda)$ is $(\Gamma + \alpha B)$-Lipschitz in t over the bounded domain considered here, the same Rademacher-contraction argument used for Assumption 1 gives a rate of order

$$\zeta_\alpha(N, \delta) = O\big((\Gamma + \alpha B)\zeta(N, \delta)\big). \tag{105}$$

Moreover, ψ_α is Lipschitz in the multiplier with constant

$$\overline{B}_\alpha := \max\left\{B, \frac{\Gamma}{\alpha}\right\}. \tag{106}$$

Lemma 4 (Augmented dual learnability)**.** *Let Assumptions 1 and 2 hold, with ν_Φ defined with respect to an optimal augmented multiplier $\lambda_\alpha^\star$. Then, with probability at least $1 - 3\delta$,*

$$\left|D_\alpha^\star - \widehat{D}_{\Phi,\alpha}^\star\right| \leq \overline{B}_\alpha \nu_\Phi + \zeta_0(N_0, \delta) + \zeta_\alpha(N, \delta) =: \Delta_D^\alpha. \tag{107}$$

Lemma 4 is the direct augmented counterpart of Lemma 1. The only changes are the Lipschitz constant $\overline{B}_\alpha$ associated with the PHB penalty and the corresponding uniform-convergence rate ζ_α.

We next state the primal recovery result. Define

$$C_\alpha := 1 + \Gamma + \alpha B \tag{108}$$

and

$$\varepsilon_\alpha := \Delta_D^\alpha + \zeta_0 + L C_\alpha \nu_{\mathcal{H}}. \tag{109}$$

Further define

$$\Delta_R^\alpha := \sqrt{\frac{2}{\alpha}\varepsilon_\alpha} + \sqrt{\frac{2L^3 C_\alpha \nu_{\mathcal{H}}}{\mu}}. \tag{110}$$

Theorem 3 (Statistical primal recovery with augmentation)**.** *Let the assumptions of Theorem 1 hold, and let*

$$\widehat{f}_{\theta_{\Phi,\alpha}} \in \arg\min_{f_\theta \in \mathcal{H}} \widehat{\mathcal{L}}_\alpha \left(f_\theta, \widehat{\lambda}^\star_{\Phi,\alpha} \right). \tag{111}$$

Then, with probability at least $1 - 4\delta$*,*

$$\mathbb{E}_{z\sim\mathcal{D}} \left[[\ell(\widehat{f}_{\theta_{\Phi,\alpha}}, z)]_+ \right] \leq \Delta^\alpha_R + \zeta(N, \delta) =: \Delta^\alpha_F, \tag{112}$$

$$\left| \mathbb{E}_{\mathcal{D}_0} \left[\ell_0(\widehat{f}_{\theta_{\Phi,\alpha}}, z) \right] - P^\star \right| \leq \zeta_0 + \Delta^\alpha_D + L \left(1 + \frac{B_0}{s} \right) \nu_{\mathcal{H}} + \left(\Gamma + \frac{\alpha}{2} \max \left\{ B, \frac{\Gamma}{\alpha} \right\} \right) \Delta^\alpha_R. \tag{113}$$

The sensitivity result extends in the same way. We additionally assume, as standard for the PHB formulation, that an optimal augmented multiplier is also a Lagrange multiplier of the original constrained problem.

Theorem 4 (Empirical approximate sensitivity with augmentation)**.** *Let the assumptions of Theorem 3 and Assumption 7 hold. Then, with probability at least* $1 - 4\delta$*, simultaneously for every admissible perturbation* u*,*

$$P(u) \geq P(0) - \mathbb{E}_{z\sim\mathcal{D}} \left[\widehat{\lambda}^\star_{\Phi,\alpha}(z) u(z) \right] - \left[L \left(1 + \frac{B_0}{s} \right) \nu_{\mathcal{H}} + \overline{B}_\alpha \nu_\Phi + 2\zeta_0 + 2\zeta_\alpha + \overline{B}_\alpha \zeta_\Lambda \right]. \tag{114}$$

Comparison with the standard Lagrangian. The augmented and non-augmented guarantees have the same qualitative structure: the recovery error is determined by statistical estimation, approximation of the optimal multiplier by Λ_Φ, and approximation of the convexified predictor by the original hypothesis class. The main difference appears in the conversion from dual suboptimality to primal feasibility.

For the standard Lagrangian, strong convexity of the objective implies that the convexified dual has gradient Lipschitz constant L^2/μ, which produces the term

$$\sqrt{\frac{2L^2}{\mu}\varepsilon}. \tag{115}$$

In contrast, the PHB augmented dual is $1/\alpha$-smooth. The corresponding term is therefore

$$\sqrt{\frac{2}{\alpha}\varepsilon_\alpha}. \tag{116}$$

Thus, augmentation replaces the dependence on the curvature of the primal objective by a dependence on the augmentation parameter α. In particular, larger α directly improves the conversion of dual suboptimality into constraint satisfaction. When $\nu_{\mathcal{H}} = 0$, this removes the dependence on μ from the feasibility argument entirely.

There is nevertheless a tradeoff: increasing α also increases the Lipschitz constant of the augmented loss, and therefore the terms C_α and ζ_α. Hence, α improves the conditioning of the dual recovery step while potentially increasing approximation and statistical errors associated with the augmented objective.

A.5.1 PROOFS

Proof of Lemma 4. The proof follows that of Lemma 1. The only modification is the comparison between two multipliers. From equation 99,

$$\left| \frac{\partial}{\partial\lambda} \psi_\alpha(t, \lambda) \right| = \left| \frac{[\lambda + \alpha t]_+ - \lambda}{\alpha} \right| \leq \overline{B}_\alpha. \tag{117}$$

Hence,

$$|\psi_\alpha(\ell(f_\theta, z), \lambda_\phi(z)) - \psi_\alpha(\ell(f_\theta, z), \lambda^\star_\alpha(z))| \leq \overline{B}_\alpha |\lambda_\phi(z) - \lambda^\star_\alpha(z)|. \tag{118}$$

Integrating gives the term $\overline{B}_\alpha \nu_\Phi$. The population-to-empirical comparison is identical after replacing the weighted constraint-loss uniform-convergence term $\Gamma\zeta$ by ζ_α. The remainder of the proof is unchanged.

Proof of Theorem 3. The proof follows the convexification argument of Theorem 1. The key modification is the dual-smoothness step. Define the PHB residual

$$r_\alpha(f,\lambda) := \frac{[\lambda + \alpha\ell(f)]_+ - \lambda}{\alpha}. \tag{119}$$

For the convexified empirical augmented dual,

$$\nabla \widehat{d}_{C,\alpha}(\lambda) = r_\alpha(\widehat{f}_{C,\alpha}(\lambda),\lambda). \tag{120}$$

The PHB augmented dual is $1/\alpha$-smooth. Therefore, the cocoercivity argument used in the proof of Theorem 1 now gives

$$\left| r_\alpha(\widehat{f}_{C,\alpha}(\widehat{\lambda}^\star_{\Phi,\alpha}),\widehat{\lambda}^\star_{\Phi,\alpha}) \right|_{2,N} \le \sqrt{\frac{2}{\alpha}\varepsilon_\alpha}. \tag{121}$$

Moreover, coordinatewise,

$$[\ell(f,z)]_+ = [r_\alpha(f,\lambda)(z)]_+, \tag{122}$$

so equation 121 directly controls empirical constraint violation.

The transfer from $\mathrm{conv}(\mathcal{H})$ to $\mathcal{H}$ proceeds exactly as in Theorem 1. The only change is that the augmented Lagrangian is LC_α-Lipschitz with respect to the approximation error, replacing the factor $L(1+\Gamma)$. Strong convexity then yields the second term in equation 110. Uniform convergence of the positive-part constraint loss completes equation 112.

For optimality, use the identity

$$\psi_\alpha(t,\lambda) = \lambda r + \frac{\alpha}{2}r^2, \qquad r = \frac{[\lambda+\alpha t]_+ - \lambda}{\alpha}. \tag{123}$$

The augmented contribution is therefore controlled by the residual bound equation 110. The remaining population–empirical and convexification terms are identical to those in the proof of Theorem 1.

Proof of Theorem 4. The sensitivity argument is unchanged because an optimal PHB multiplier is also a multiplier of the original constrained problem and therefore has the same sensitivity interpretation. The proof of Theorem 2 can consequently be repeated verbatim after replacing Lemma 1 by Lemma 4. In particular, the terms $B\nu_\Phi + \Gamma\zeta$ appearing in the standard-dual argument are replaced by $\overline{B}_\alpha\nu_\Phi + \zeta_\alpha$. The dual-class generalization term $B\zeta_\Lambda$ is unchanged.

B Tasks Details

B.1 MNIST image classification

Problem formulation. We first consider constrained image classification on MNIST. Rather than minimizing a conventional classification loss, the objective penalizes only the norm of the model parameters,

$$f(\theta) = \frac{c}{2}\|\theta\|_2^2, \tag{124}$$

while correct classification is imposed through pointwise margin constraints.

Let $z_j(x_i;\theta)$ denote the logit assigned to class j for image x_i, and let y_i denote its ground-truth class. For every incorrect class $j \neq y_i$, we impose

$$g_{ij}(\theta) := \epsilon - (z_{y_i}(x_i;\theta) - z_j(x_i;\theta)) \le 0. \tag{125}$$

Thus, feasibility requires the correct-class logit to exceed every incorrect-class logit by at least the prescribed margin ϵ. We use

$$\epsilon = 1, \qquad c = 0.1. \tag{126}$$

MNIST contains

$$N = 60{,}000 \tag{127}$$

training images. Since each image generates one constraint for each of its nine incorrect classes, the total number of pointwise constraints is

$$M = 60{,}000 \times 9 = 540{,}000. \tag{128}$$

Primal model. The classifier is a convolutional neural network composed of two 3×3 convolutional layers of width 64, followed by ReLU activations and max pooling, a fully connected hidden layer of width 32, and a linear output classifier. The resulting primal network contains

$$p = 138{,}282 \tag{129}$$

trainable parameters. Batch normalization is not used.

Dual parameterization. For MNIST, the parametric multiplier is implemented as a second convolutional network with the same architecture as the primal classifier. In particular, the dual network consists of two 3×3 convolutional layers of width 64 with ReLU activations and max pooling, followed by a fully connected hidden layer of width 32 and a linear output layer. Given an image x_i, the network produces one multiplier for each class,

$$h_\phi(x_i) \in \mathbb{R}^{10}. \tag{130}$$

The outputs are mapped to nonnegative bounded multipliers according to

$$\lambda_\phi(x_i) = \Gamma\, \sigma(h_\phi(x_i))\,, \tag{131}$$

where Γ is the multiplier cap and σ is the logistic sigmoid. Since the true class does not define a margin constraint, its corresponding component is fixed to zero,

$$[\lambda_\phi(x_i)]_{y_i} = 0. \tag{132}$$

The output-layer bias is initialized to -7, so that the initial multiplier predictions are close to zero.

B.2 FLORA-BENCH WORKFLOW PREDICTION

Problem formulation. We next consider the Coding-GD split of FLORA-Bench, where a graph neural network predicts whether a multi-agent workflow successfully solves a coding task. Each observation corresponds to a workflow–task pair.

The objective encourages successful workflows to rank above unsuccessful workflows within each task. Let $\mathcal{P}_t$ and $\mathcal{N}_t$ denote the successful and unsuccessful workflows associated with task t, respectively. We use the pairwise ranking objective

$$f(\theta) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{1}{|\mathcal{P}_t||\mathcal{N}_t|} \sum_{i \in \mathcal{P}_t} \sum_{j \in \mathcal{N}_t} \log\left(1 + \exp\left(-\left[s_i(\theta) - s_j(\theta)\right]\right)\right), \tag{133}$$

where $s_i(\theta)$ denotes the predicted logit for workflow–task pair i.

The pointwise constraints additionally require every individual pair to be classified with sufficiently small binary cross-entropy,

$$g_i(\theta) := \mathrm{BCE}_i(\theta) - \tau \le 0, \tag{134}$$

with

$$\tau = 0.1. \tag{135}$$

The training set contains

$$N = 24{,}489 \tag{136}$$

observed workflow–task pairs.

Dataset structure. The data form a partially observed task–workflow matrix with

$$N_{\text{task}} = 57, \qquad N_{\text{workflow}} = 1026. \tag{137}$$

The complete grid therefore contains

$$57 \times 1026 = 58{,}482 \tag{138}$$

possible pairs, of which approximately

$$\frac{24{,}489}{58{,}482} \approx 0.419 \tag{139}$$

are observed. Each workflow appears under multiple tasks,

$$14 \le N_{\text{tasks/workflow}} \le 51, \tag{140}$$

and each task contains between

$$221 \le N_{\text{workflows/task}} \le 616 \tag{141}$$

workflows.

Primal model. The primal predictor is a graph neural network with

$$p = 26{,}755{,}073 \tag{142}$$

trainable parameters. The primal architecture is held fixed across all experiments that vary the capacity of the dual model.

Dual parameterization. Unlike MNIST, the dual predictor does not operate directly on the raw workflow graph. Instead, it is conditioned on the penultimate representation produced by the primal GNN. Let

$$h_\theta(w, t) \in \mathbb{R}^{d_h} \tag{143}$$

denote the primal representation associated with workflow w and task t. The multiplier predictor takes this representation as input,

$$\lambda_\phi(w, t) = q_\phi\left(h_\theta(w, t)\right). \tag{144}$$

Since each observed workflow–task pair has a single BCE constraint, the dual output is scalar,

$$\lambda_\phi(w, t) \in \mathbb{R}_+. \tag{145}$$

The same primal representation is used throughout the capacity experiments, so changing the dual-network width modifies only the expressivity of the multiplier class and not the primal model or its features.

B.3 TOFU LANGUAGE-MODEL UNLEARNING

Problem formulation. We next study constrained language-model unlearning using the TOFU benchmark. The starting model is Llama-3.2-1B, previously finetuned on

$$N_{\mathrm{FT}} = 4{,}000 \tag{146}$$

question–answer pairs describing fictitious authors. The forget set contains

$$N_{\mathrm{forget}} = 400 \tag{147}$$

question–answer pairs associated with

$$N_{\mathrm{authors}} = 20 \tag{148}$$

authors.

The primal objective preserves performance on a fixed retain set by minimizing its negative log-likelihood,

$$f(\theta) = \frac{1}{|\mathcal{R}|} \sum_{(x,y)\in\mathcal{R}} -\log p_\theta(y \mid x), \tag{149}$$

using a retain set of

$$|\mathcal{R}| = 256 \tag{150}$$

sequences.

For each forget example i, we impose four pointwise constraints. Let $s_i(\theta)$ denote the length-normalized answer log-probability ratio relative to the reference model. The first two constraints impose upper and lower bounds on this score,

$$s_i(\theta) \le u_i, \qquad s_i(\theta) \ge l_i, \tag{151}$$

where the bounds are constructed from a per-example target with the prescribed shift and band. In addition, two constraints control the token-level log-probability profile by imposing an upper bound on the average log-probability of the most likely answer tokens and a lower bound on the average log-probability of the least likely answer tokens. Thus, each forget example contributes four scalar pointwise constraints. The reported experiments use a shift and band of 0.25 and an optimization budget of 20 epochs.

Dual parameterization. Each of the four constraints associated with a forget example has its own nonnegative multiplier. The parametric formulation replaces the independent tabular multipliers with a shared dual network,

$$\lambda_\phi(x_i) \in \mathbb{R}^4_+. \tag{152}$$

The dual network takes a pooled hidden representation of the forget example and outputs one multiplier for each constraint,

$$\lambda_\phi(x_i) = \left[\lambda_{\phi,1}(x_i), \lambda_{\phi,2}(x_i), \lambda_{\phi,3}(x_i), \lambda_{\phi,4}(x_i)\right]^\top. \tag{153}$$

This construction allows the learned dual to predict the corresponding multipliers for forget examples that were not used to fit the dual network.

Evaluation metrics. In addition to the retain objective and pointwise forgetting constraints, we report the standard TOFU metrics for memorization, privacy, and utility. Their aggregate is computed as a harmonic mean,

$$S_{\text{agg}} = \text{HMean}\left(S_{\text{mem}}, S_{\text{priv}}, S_{\text{util}}\right). \tag{154}$$

Reporting these components separately is important because similar aggregate scores can result from qualitatively different unlearning behaviors.

B.4 IEEE30 Optimal Power flow dispatch

Problem formulation. Optimal power flow (OPF) is the central optimization problem underlying economic dispatch in electrical grids. It is repeatedly solved to allocate generation, determine operating points, and support electricity-market and infrastructure decisions Cain et al. (2012). We represent the grid as a network of N buses connected by a set $\mathcal{E}$ of M directed branches Chatzivasileiadis (2018). Each bus i is associated with complex voltage v_i, complex generation s_i, and demand r_i.

Under the standard π-section branch model Coffrin et al. (2018); Zimmerman et al. (2011), the complex power flows through branch (i, j) are

$$\begin{aligned} f_{ij} &= \left(y_{ij} + y_{ij}^C\right)^* \left|\frac{v_i}{t_{ij}}\right|^2 - y_{ij}^* \frac{v_i v_j^*}{t_{ij}}, \\ f_{ji} &= \left(y_{ij} + y_{ji}^C\right)^* |v_j|^2 - y_{ij}^* \frac{v_i^* v_j}{t_{ij}^*}, \end{aligned} \tag{155}$$

where t_{ij} denotes the transformer ratio, y_{ij} the line admittance, and y_{ij}^C and y_{ji}^C the shunt admittances.

The generation cost is quadratic in the real generated power,

$$C(s) = \sum_{i=1}^{N} c_{0i} + c_{1i}\,\text{Re}(s_i) + c_{2i}\,\text{Re}^2(s_i). \tag{156}$$

Given a demand vector

$$r = [r_1; \ldots; r_N], \tag{157}$$

the OPF problem minimizes generation cost while enforcing power balance together with generation, voltage, thermal, and phase-angle limits,

$$\begin{aligned} \min_{s,v} \quad & C(s) \\ \text{s.t.} \quad & s_i - r_i = \sum_{j \in n(i)} f_{ij} - \left(y_i^S\right)^* |v_i|^2, \\ & s_{i,\min}^G \le s_i \le s_{i,\max}^G, \\ & v_{i,\min} \le |v_i| \le v_{i,\max}, \\ & |f_{ij}| \le f_{ij,\max}, \qquad |f_{ji}| \le f_{ji,\max}, \\ & \theta_{ij,\min} \le \angle(v_i v_j^*) \le \theta_{ij,\max}, \end{aligned} \tag{158}$$

with branch flows defined by equation 155. Collecting the grid parameters into $\mathcal{W}$, we write the problem compactly as

$$\min_{s,v} \; C(s) \qquad \text{s.t.} \qquad g(s, v; r, \mathcal{W}) \leq 0, \qquad h(s, v; r, \mathcal{W}) = 0. \tag{159}$$

The AC-OPF problem is nonconvex Lesieutre & Hiskens (2005) and NP-hard in general Bienstock & Verma (2019). Classical interior-point methods and convex relaxations can provide accurate solutions Molzahn & Hiskens (2016); Low (2014); Kocuk et al. (2016), but repeatedly solving OPF as demand changes can be computationally expensive Castillo & O'Neill (2013). This motivates learning an amortized predictor

$$[s, v] = \Phi(r; \mathcal{A}, \mathcal{W}), \tag{160}$$

typically parameterized by a graph neural network exploiting the grid topology Owerko et al. (2020; 2024).

Pointwise constrained learning. Supervised approaches regress onto solutions produced by an optimization solver Guha et al. (2019); Owerko et al. (2020), whereas unsupervised methods directly optimize the OPF objective and physical constraints Owerko et al. (2024). Constraint penalties or expectation-based constrained formulations can nevertheless allow violations on individual demand realizations to be compensated by slack on others Huang et al. (2022); Pan et al. (2023); Wang et al. (2021); Huang et al. (2024); Fioretto et al. (2020); Chen et al. (2022); Chamon & Ribeiro (2020).

Following Owerko et al. (2025), we instead consider the pointwise-constrained learning problem

$$\begin{aligned} \min_{\mathcal{A}} \quad & \mathbb{E}_{r\sim\rho}\left[C\left(\Phi(r; \mathcal{A}, \mathcal{W})\right)\right] \\ \text{s.t.} \quad & g\left(\Phi(r; \mathcal{A}, \mathcal{W}); r, \mathcal{W}\right) \leq 0, \\ & h\left(\Phi(r; \mathcal{A}, \mathcal{W}); r, \mathcal{W}\right) = 0, \qquad \forall r. \end{aligned} \tag{161}$$

This formulation associates each demand realization with its own inequality and equality multipliers and can be solved through an augmented-Lagrangian primal–dual procedure Wierzbicki & Kurcyusz (1977); Owerko et al. (2025). The resulting multipliers additionally admit a sensitivity interpretation, quantifying how strongly the optimal dispatch cost depends on the corresponding physical constraints Boyd & Vandenberghe (2004).

Dataset and primal model. We evaluate on the IEEE 30-bus system, which contains

$$N_{\text{bus}} = 30, \qquad N_{\text{gen}} = 6, \qquad N_{\text{branch}} = 41. \tag{162}$$

The dataset contains

$$N_{\text{sample}} = 8{,}000 \tag{163}$$

demand realizations. The primal predictor is the graph neural network used for pointwise OPF learning in Owerko et al. (2025), which maps each demand realization to the corresponding generation and voltage variables.

Dual structure. In the pointwise formulation, every demand realization has an independent structured vector of multipliers. The complete tabular representation contains five entries per bus, four per generator, and six per branch, resulting in

$$M_{\text{tab}} = 5N_{\text{bus}} + 4N_{\text{gen}} + 6N_{\text{branch}} = 420 \tag{164}$$

multipliers per operating condition.

The first

$$M_{\text{eq}} = 90 \tag{165}$$

entries correspond to equality constraints and are therefore free-sign, while the remaining entries correspond to inequality constraints and are nonnegative. Storing one such vector independently for every training realization causes the number of dual variables to grow linearly with the size of the dataset Park & Van Hentenryck (2023); Owerko et al. (2025).

Dual parameterization. We replace this per-realization multiplier table with a shared dual network. Rather than maintaining an independent multiplier vector for every demand profile, the parametric dual reads the latent node representations produced by the primal GNN and predicts the corresponding multipliers for buses, generators, and transmission branches.

Let

$$h_\theta(r) \in \mathbb{R}^{d_h} \tag{166}$$

denote the penultimate representation of the primal GNN for demand realization r. The parameterized multiplier is obtained as

$$\lambda_\phi(r) = q_\phi\left(h_\theta(r)\right). \tag{167}$$

The output preserves the physical grouping of the OPF constraints, so that separate predictions are produced for the multiplier components associated with buses, generators, and branches.

The parametric implementation outputs

$$M_{\text{param}} = 338 \tag{168}$$

multipliers per operating condition. The first

$$90 \tag{169}$$

components correspond to free-sign equality multipliers, while the remaining outputs are constrained to be nonnegative. The parametric representation omits one two-dimensional branch constraint group corresponding to voltage-angle differences. Consequently, comparisons between the tabular and parametric duals are performed on their common inequality subset,

$$\mathcal{I} = \{90, \dots, 337\}, \tag{170}$$

which contains

$$|\mathcal{I}| = 248 \tag{171}$$

inequality multipliers per operating condition.

This construction changes the scaling of the dual representation. Whereas the tabular formulation stores a distinct multiplier vector for every training sample, the parametric formulation stores only the parameters ϕ of the shared dual network. More importantly for our setting, it defines

$$\lambda_\phi(r) \tag{172}$$

for previously unseen demand realizations, allowing sample-level constraint sensitivities to be evaluated beyond the training set.

C Implementation Details

We now describe how the parametric dual problem is optimized in practice. Given the constraint samples

$$S = \{z_n\}_{n=1}^N,$$

we write the empirical constraints as

$$g_n(\theta) := \ell(f_\theta, z_n), \qquad n = 1, \dots, N, \tag{173}$$

where the constraint threshold has been absorbed into the definition of ℓ, consistently with the formulation in Section 2.

For a constraint value g, multiplier $\lambda \geq 0$, and penalty parameter $\rho > 0$, we use the Powell–Hestenes–Rockafellar augmented Lagrangian term

$$\psi_\rho(g, \lambda) := \frac{[\lambda + \rho g]_+^2 - \lambda^2}{2\rho}. \tag{174}$$

The empirical augmented Lagrangian associated with a parametric multiplier $\lambda_\phi \in \Lambda_\Phi$ is therefore

$$\widehat{L}_\rho(f_\theta, \lambda_\phi) := \frac{1}{N_0} \sum_{z_n \in S_0} \ell_0(f_\theta, z_n) + \frac{1}{N} \sum_{z_n \in S} \psi_\rho\big(\ell(f_\theta, z_n), \lambda_\phi(z_n)\big). \tag{175}$$

Pointwise and parametric dual updates. The standard pointwise formulation maintains an independent multiplier λ_n for every constraint and performs the projected ascent update

$$\lambda_n^+ = \Pi_{[0,\Gamma]}\left(\lambda_n + \rho\,\ell(f_\theta, z_n)\right), \tag{176}$$

where Γ is the multiplier bound introduced in Section 3.

The parametric formulation replaces the collection $\{\lambda_n\}_{n=1}^N$ with a single function $\lambda_\phi \in \Lambda_\Phi$. At outer iteration k, we first evaluate the current multiplier function on the training constraints,

$$\lambda_{\phi_k}(z_n), \tag{177}$$

and form the corresponding pointwise augmented-Lagrangian ascent targets,

$$\widetilde{\lambda}_n^{k+1} = \Pi_{[0,\Gamma]}\left(\lambda_{\phi_k}(z_n) + \rho\,\ell(f_{\theta_k}, z_n)\right). \tag{178}$$

When smoothing is used, we replace this target by

$$\bar{\lambda}_n^{k+1} = \beta\lambda_{\phi_k}(z_n) + (1-\beta)\widetilde{\lambda}_n^{k+1}, \qquad 0 \le \beta < 1. \tag{179}$$

The updated dual function is then obtained by approximately projecting these pointwise ascent values back onto the parametric family,

$$\phi_{k+1} \approx \arg\min_\phi \frac{1}{|\mathcal{I}_{\text{fit}}|} \sum_{n\in\mathcal{I}_{\text{fit}}} \left|\lambda_\phi(z_n) - \bar{\lambda}_n^{k+1}\right|^2. \tag{180}$$

Thus, the regression step is not used to fit a separately optimized pointwise multiplier table. Rather, it implements an approximate projection of an augmented-Lagrangian dual-ascent step back onto Λ_Φ. No persistent sample-wise multipliers are maintained: the dual state of the parametric method is entirely represented by ϕ.

Primal update. For fixed ϕ_k, the primal parameters are updated by approximately minimizing the empirical augmented Lagrangian

$$\theta_{k+1} \approx \arg\min_\theta \widehat{L}_\rho(f_\theta, \lambda_{\phi_k}), \tag{181}$$

using stochastic gradient descent. Equivalently, for a minibatch $\mathcal{B} \subseteq S$, the constraint contribution is

$$\frac{1}{|\mathcal{B}|}\sum_{z_n\in\mathcal{B}} \frac{[\lambda_{\phi_k}(z_n) + \rho\,\ell(f_\theta, z_n)]_+^2 - \lambda_{\phi_k}(z_n)^2}{2\rho}. \tag{182}$$

Primal and dual updates are alternated throughout training.

Relation to the parametric dual problem. Equations 178–180 can be interpreted as a practical approximate ascent procedure for the parametric empirical dual problem. The first step performs the standard augmented-Lagrangian ascent update at the observed constraints, while the second projects the resulting multiplier values back onto the shared function class Λ_Φ. The projection is approximate because λ_ϕ is represented by a neural network and equation 180 is solved by a finite number of stochastic-gradient steps. Consequently, the theoretical results, which are stated for an exact optimizer of the parametric empirical dual, should be understood as characterizing the limiting optimization problem; the implementation provides its stochastic approximate realization.

Held-out evaluation of the dual function. To evaluate whether λ_ϕ learns a transferable multiplier rule rather than memorizing the observed constraints, we optionally partition the constraint samples into a dual-fitting set $\mathcal{I}_{\text{fit}}$ and a held-out set $\mathcal{I}_{\text{test}}$. The primal optimization continues to use all constraints, whereas equation 180 is optimized only over $\mathcal{I}_{\text{fit}}$. Multiplier predictions on $\mathcal{I}_{\text{test}}$ therefore provide an out-of-sample evaluation of the learned dual function.

For evaluation only, we compare these predictions against multipliers obtained from an independently trained pointwise-dual reference. This reference is not used as a regression target for the parametric method. .

Algorithm 1 Parametric augmented-Lagrangian optimization

Require: Objective sample S_0, constraint sample $S = \{z_n\}_{n=1}^{N}$, primal parameters θ, dual parameters ϕ, penalty ρ, multiplier bound Γ, smoothing parameter β, dual-fitting set $\mathcal{I}_{\mathrm{fit}}$
Ensure: Trained primal predictor f_θ and parametric multiplier λ_ϕ

```
Initialize θ and φ
Optionally warm up the primal parameters
for k = 1, 2, ... do
    Primal update:
    for each minibatch B ⊆ S do
        Evaluate λ_{φ_k}(z_n) for z_n ∈ B
        Update θ by descending a stochastic estimate of
        L̂_ρ(f_θ, λ_{φ_k})
    end for
    Dual ascent targets:
    for n ∈ I_fit do
        λ̃_n^{k+1} ← Π_{[0,Γ]} ( λ_{φ_k}(z_n) + ρ ℓ(f_{θ_{k+1}}, z_n) )
        λ̄_n^{k+1} ← β λ_{φ_k}(z_n) + (1 − β) λ̃_n^{k+1}
    end for
    Projection onto Λ_Φ:
    for j = 1, ..., T_dual do
        Sample B_λ ⊆ I_fit
        Update φ by descending
        (1 / |B_λ|) Σ_{n∈B_λ} | λ_φ(z_n) − λ̄_n^{k+1} |^2
    end for
    Optionally update ρ according to the constraint-violation schedule
end for
```

D Additional Results

This section provides additional experimental results complementing those presented in the main text. We first report primal-recovery results for MNIST, FLORA-Bench, and a disaggregated analysis of the IEEE 30-bus experiment. We then provide additional capacity and sensitivity studies on MNIST and FLORA-Bench, respectively. We focus these more computationally intensive ablations on MNIST and FLORA-Bench because they permit repeated training and evaluation across multiple dual architectures and perturbation experiments at substantially lower computational cost. Together, these experiments further support the two properties studied in the main text: that parametric duals preserve the quality of the recovered primal solution and that their predictions retain meaningful information about sample-level sensitivity on unseen data.

D.1 Additional Primal Recovery Results

Primal recovery on MNIST. Figure 4 complements the language-model unlearning experiment in the main text with primal-recovery results on MNIST. The pointwise and parameterized dual formulations produce nearly identical constraint profiles, driving the constraint values close to the feasible region across almost the entire test distribution. At the same time, the two formulations attain comparable primal objective values and essentially identical classification performance. These results show that replacing independent sample-wise multipliers with a shared parametric function $\lambda_\phi(z)$ does not substantially alter the recovered primal model on MNIST.

Primal recovery on FLORA-Bench. A similar behavior is observed on FLORA-Bench, as shown in Figure 5. Both the pointwise and parameterized dual formulations drive the pointwise constraint violations close to zero and attain the same primal objective value, while maintaining essentially identical test accuracy. In contrast, ERM and feasibility-only training obtain substantially larger objective values and exhibit worse constraint satisfaction. Thus the parameterized formulation preserves the quality of the primal solution.

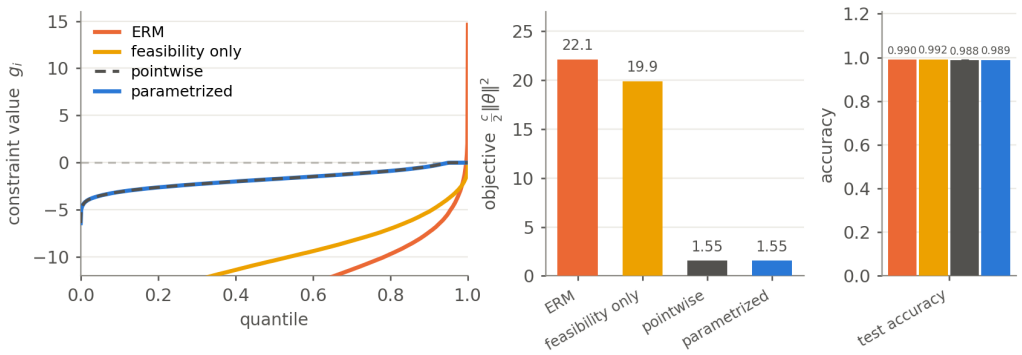


Figure 4: **Primal recovery on MNIST.** Comparison of the parameterized dual formulation with independent pointwise multipliers and the corresponding unconstrained or partially constrained baselines. The parameterized formulation closely matches the pointwise solution in constraint satisfaction and primal objective while preserving classification performance.

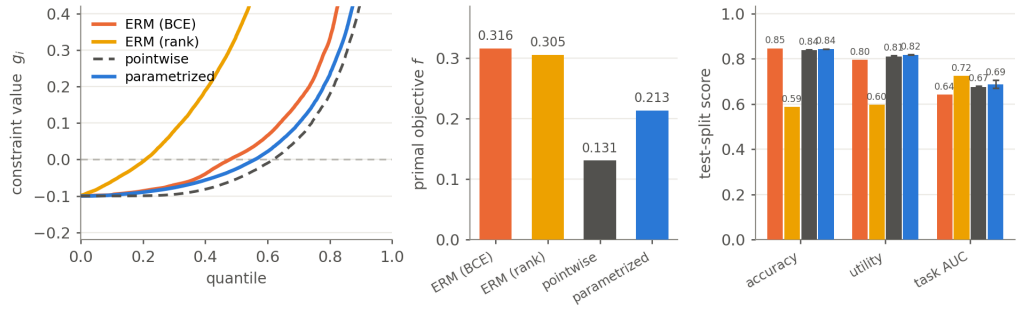


Figure 5: **Primal recovery on FLORA-Bench.** Comparison of ERM, feasibility-only training, independent pointwise multipliers, and parameterized dual multipliers. Pointwise and parameterized duals achieve near-zero constraint violations and comparable objective values while maintaining essentially identical test accuracy.

Constraint violations by type on IEEE 30-bus. Table 6 complements the aggregate IEEE 30-bus results reported in the main text by disaggregating violations according to constraint type. The parameterized formulation maintains small violations across the different physical quantities, despite learning a shared representation for hundreds of heterogeneous constraints. Its error profile differs across constraint classes: the largest deviations arise for reactive generation and branch-flow constraints, whereas active generation and voltage-magnitude violations remain small. This finer-grained analysis reinforces the aggregate result from the main text: the learned dual parameterization remains effective even when simultaneously representing constraints that act on different variables and different parts of the network.

D.2 Additional Dual Capacity Results

Dual capacity on MNIST. We further study how the capacity of the dual model affects recovery of the underlying pointwise multipliers on MNIST. Figure 7 varies the number of parameters in λ_ϕ over several orders of magnitude and reports both ranking recovery and the corresponding approximation–generalization decomposition. Increasing dual capacity improves ranking recovery, particularly for the smaller parameterizations, indicating that insufficiently expressive networks are primarily limited by approximation error. As capacity increases further, the gains become progressively smaller and eventually saturate. In contrast to the stronger train–test tradeoff observed in the Flora-Bench experiment, the generalization error on MNIST remains comparatively stable over the range of model sizes considered. This behavior is likely facilitated by the large number of samples available in MNIST, which provides substantially more data for learning increasingly expressive dual functions. This experiment illustrates that the tradeoff between dual expressivity and statistical

	$\mathrm{Re}(\mathbf{s}^g)$ (%)		$\mathrm{Im}(\mathbf{s}^g)$ (%)		$\lvert V\rvert$ (%)		$\lvert\mathbf{f}_f\rvert$ (%)		$\lvert\mathbf{f}_t\rvert$ (%)	
	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max
Param	0.32	0.53	5.68	5.68	0.12	0.12	4.90	4.90	2.49	2.49
Dual-P	0.20	0.25	0.69	0.70	0.74	0.85	0.68	0.68	0.15	0.15
Dual-H	0.35	0.72	0.90	0.92	3.76	5.61	0.25	0.25	0.00	0.00
Dual-S	0.55	0.86	0.14	0.14	6.00	8.21	0.00	0.00	0.00	0.00

Figure 6: **Constraint violations on IEEE 30-bus, disaggregated by constraint type.** For each test operating point, the mean and maximum violation are computed separately for active generation $\mathrm{Re}(\mathbf{s}^g)$, reactive generation $\mathrm{Im}(\mathbf{s}^g)$, voltage magnitude $|V|$, and branch-flow magnitudes $|\mathbf{f}_f|$ and $|\mathbf{f}_t|$, and are then averaged across the test set. Violations are expressed as percentages of the corresponding feasible range. Voltage-angle constraints are omitted because no violations are observed. Baseline values are reproduced from Owerko et al. (2025), Fig. 11.

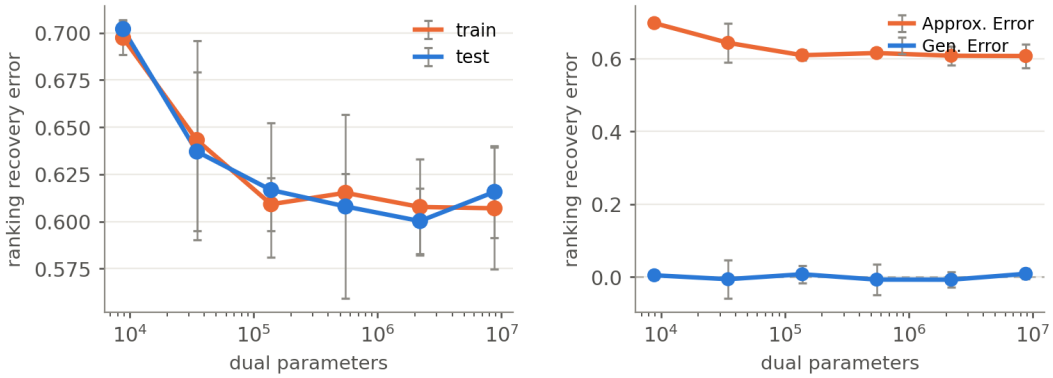


Figure 7: **Effect of dual parameterization capacity on MNIST.** Ranking recovery and error decomposition as the number of parameters in the dual network increases. Increasing capacity primarily reduces approximation error and improves recovery of the pointwise multiplier ranking, while the generalization error remains comparatively stable across model sizes.

complexity depends jointly on the capacity of Λ_Φ and on the amount of data available to estimate the multiplier function.

D.3 Additional Sensitivity Results

Sensitivity on FLORA-Bench. Finally, we repeat the direct sensitivity experiment on FLORA-Bench. Figure 8 evaluates how the primal objective changes when constraints associated with unseen samples are imposed, grouping the samples according to the magnitude of their predicted multipliers. As the multiplier norm increases, the resulting increase in the optimized objective also tends to grow. The separation is less sharp than in the MNIST experiment reported in the main text, but the overall trend is similar: constraints assigned larger dual values tend to have a greater effect on the primal optimum when enforced. This weaker relationship is consistent with the more heterogeneous structure of FLORA-Bench, where sensitivity can depend jointly on the workflow, the task, and its interaction with the remaining constrained samples. Nevertheless, the observed increase provides direct evidence that the predicted multiplier retains meaningful information about the cost of enforcing previously unseen constraints. Together with the ranking results of Table 2, this indicates that the learned dual captures both which unseen constraints are important and, to a meaningful extent, how strongly they influence the optimized objective.

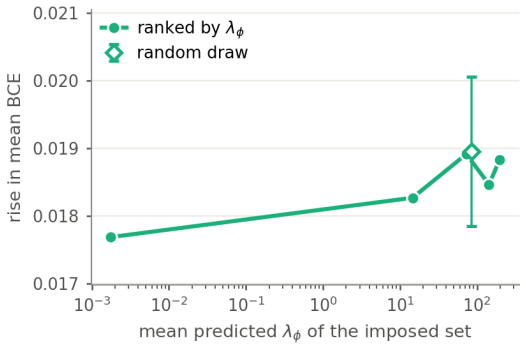


Figure 8: **Sensitivity estimation on FLORA-Bench.** Change in the optimized objective after imposing constraints on unseen samples, grouped according to the magnitude of their predicted multipliers. Sets with larger predicted multiplier values induce larger objective changes, providing an operational validation of the sensitivity information encoded by the learned dual.