Constraint-Aware Influence Estimation in Deep Constrained Learning

Xin Wang
Department of Civil & Environmental Engineering
University of Washington
Seattle, WA 98195, USA
xinw22@uw.edu

R. Tyrrell Rockafellar
Department of Mathematics
University of Washington
Seattle, WA 98195, USA
rtr@math.washington.edu

Xuegang (Jeff) Ban*
Department of Civil & Environmental Engineering
University of Washington
Seattle, WA 98195, USA
banx@uw.edu

Abstract

Constrained learning has been increasingly applied to various domains to ensure explicit feasibility requirements due to fairness, safety, robustness, regularization, and physics or logic constraints. Understanding how training samples influence the solution (e.g., learned parameters) of constrained learning is crucial for interpretability and robustness. The classical influence function may become unreliable in constrained settings: data perturbations can reshape both the objective and the feasible region, leading to estimates that violate feasibility. In response, we propose the Directional Influence Function (DIF), a new estimator that explicitly incorporates the constraints into influence estimation. DIF formulates the optimality conditions of constrained learning as a variational inequality (VI) and analyzes how perturbing training data affects this VI. We validate DIF in constrained linear regression and demonstrate that it recovers leave-one-out retraining results, whereas IF and penalty-based IF exhibit significant bias. We further apply DIF to fairness-constrained CNNs, where DIF accurately predicts test loss changes under data removal and aligns closely with actual retraining. Our results establish DIF as an efficient and reliable tool for data attribution in constrained learning.

1 Introduction

Understanding how individual training samples influence model predictions, also known as data attribution [34, 11, 28], is fundamental for interpreting model behavior, model correction, and data error debugging. Although retraining the model after removing a sample and comparing the resulting change in model solution (learned parameters) provides ground-truth data influence, this approach is computationally expensive for modern deep learning models. A widely adopted alternative is to approximate sample influence by studying how the solution responds to small data perturbations. As a representative method, the influence function (IF), originally developed in robust statistics [17] and later adapted to machine learning [23, 11, 49], instantiates this idea by differentiating the solution with respect to (w.r.t.) small data perturbations. While these influence estimators work for unconstrained learning problems, they often fail when applied to *constrained learning*. The latter refers to learning tasks constrained by domain knowledge, such as physics-informed neural networks (PINNs), or those

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

constrained by embedding requirements, including safety, robustness, and fairness [19, 27, 12, 47]. The empirical formulation of constrained learning is given by [6]:

$$\begin{aligned} \bar{\theta} := \arg\min_{\theta \in \mathbb{R}^d} \quad & \frac{1}{N_0} \sum_{i=1}^{N_0} \ell_0 \left(z_i^{(0)}, \theta \right) \\ \text{s.t.} \quad & \frac{1}{N_j} \sum_{i=1}^{N_j} \ell_j \left(z_i^{(j)}, \theta \right) \leq \tau_j, \quad j = 1, \ldots, m, \end{aligned} \tag{1}$$

where $\theta \in \mathbb{R}^d$ denotes the model parameters to be optimized, and $\bar{\theta}$ is the optimal solution. The function $\ell_0 \left(z_i^{(0)}, \theta \right)$ represents the primary loss evaluated on the training data point $z_i^{(0)}$. Each constraint $j \in \{1, \ldots, m\}$ is represented by an auxiliary loss function $\ell_j \left(z_i^{(j)}, \theta \right)$ evaluated over a separate dataset $\left\{ z_i^{(j)} \right\}_{i=1}^{N_j}$. The constant τ_j specifies the upper bound allowed for the jth constraint. The influence estimators aim to quantify the change in solution, denoted by $\Delta\theta$, when specific training data points participating in the objective function or constraints (e.g., $z_1^{(1)}$) are removed. Current IF-based approaches fail in constrained learning for two primary reasons. First, constrained problems (1) require that the perturbed solution, $\bar{\theta} + \Delta\theta$, remains within the feasible region, while IF methods ignore the constraints, not to mention that the feasible region itself may also be altered by data perturbation. Existing IF cannot guarantee that the estimated solution change is feasible. Second, IF primarily relies on the gradient of the optimal solution $\bar{\theta}$ w.r.t. data perturbations. In constrained learning, the solution can be only directionally differentiable. The full derivatives do not necessarily exist, rendering IF estimators invalid. This issue arises because the solution updates must adhere to the feasible region when constraints are active, often requiring projections to maintain feasibility. Additionally, data removal can change the active constraint set, leading to sudden shifts in the optimal solution. See the problem (3) in Section 2.1 for a concrete example.

To overcome the above limitations, this paper introduces the *Directional Influence Function* (DIF), an influence function designed for constrained learning, to estimate the impact of training points in either the loss function or the constraints on the model solution. DIF estimates the change in model solution through a directional derivative approach, aiming to address the following question:

> *How does the solution of constrained learning change when data points are removed from either the objective loss function or the constraints?*

The main contributions of the paper are: (1) We formalize data attribution for constrained learning by casting the optimality conditions as a variational inequality (VI) and performing local sensitivity analysis of this VI; (2) DIF quantifies the effect of data perturbations on model solutions via directional derivatives, thereby addressing the non-smoothness of solution changes induced by constraints; and (3) DIF can be efficiently computed by solving a quadratic program (QP); when all constraints are inactive (i.e., their KKT multipliers are zero), DIF reduces to the classical IF.

The remainder of this paper is organized as follows. Section 2 formally defines our problem and illustrates the failure of IF in a constrained learning setting through a linear regression example. Section 3 introduces the proposed DIF. Section 4 analyzes the impact of data perturbations on the VI-formulated optimality conditions, introduces a QP approach to compute DIF, and presents our main theoretical results. Section 5 evaluates DIF on several learning tasks, followed by concluding remarks in Section 7. Throughout this paper, we assume that all functions involved are twice continuously differentiable (C^2). A detailed comparison with prior work is also presented in Appendix 6.

2 Problem Formulation

Let $Z^{(0)} = \{z_i^{(0)}\}_{i=1}^{N_0}$ denote the objective dataset and $Z^{(j)} = \{z_i^{(j)}\}_{i=1}^{N_j}$ denote the dataset for constraint j. We begin by examining how the solution $\bar{\theta}$ changes when a selected subset $Z^r \subset \bigcup_{j=0}^{m} Z^{(j)}$ is removed. This removal is modeled via a perturbation formulation, where each point in Z^r is assigned a small weight $\varepsilon_k : \varepsilon_0$ for points contributing to the objective, and $\varepsilon_j, j \in \{1, \ldots, m\}$ for the points contributing to the j-th constraint. This formulation, referred to as

Perturbed Constrained Learning (PCL), is defined as follows[2]:

$$\begin{aligned} \min_{\theta} \quad & \frac{1}{N_0}\sum_{i=1}^{N_0}\ell_0\left(z_i^{(0)},\theta\right)+\varepsilon_0\sum_{z_i^{(0)}\in Z^r}\ell_0\left(z_i^{(0)},\theta\right) \\ \text{s.t.} \quad & \frac{1}{N_j}\sum_{i=1}^{N_j}\ell_j\left(z_i^{(j)},\theta\right)+\varepsilon_j\sum_{z_i^{(j)}\in Z^r}\ell_j\left(z_i^{(j)},\theta\right)\le\tau_j, j=1,\dots,m. \end{aligned} \tag{2}$$

Let $\varepsilon=[\varepsilon_0,\dots,\varepsilon_m]$ denote the perturbation vector. When $\varepsilon=\mathbf{0}$, problem (2) reduces to the problem (1). Shifting ε from $\bar{\varepsilon}=[0,0,\dots,0]$ to $\hat{\varepsilon}=\left[-\frac{1}{N_0},-\frac{1}{N_1},\dots,-\frac{1}{N_m}\right]$ corresponds to removing Z^r from the constrained learning (1). Treat θ as a function of ε, denoted by $\theta(\varepsilon)$, and define the solution change as $\Delta\theta=\theta(\hat{\varepsilon})-\theta(\bar{\varepsilon})$. By definition, $\theta(\bar{\varepsilon})=\bar{\theta}$ is the solution to problem (1).

2.1 Failure of IF in Constrained Learning

Toy example. We use a ℓ_2-constrained least-squares example to demonstrate the failure of IF in constrained learning, as shown below. Here $\theta=[\theta_1,\theta_2]$ are the model parameters.

$$\begin{aligned} \min_{\theta} \quad & \frac{1}{3}(\theta^\top x_1-y_1)^2+\frac{1}{3}(\theta^\top x_2-y_2)^2+\frac{1}{3}(\theta^\top x_3-y_3)^2+\varepsilon\cdot(\theta^\top x_2-y_2)^2 \\ \text{s.t.} \quad & \|\theta\|_2^2\le 0.5. \end{aligned} \tag{3}$$

The data are as follows:$[x_1,x_2,x_3]=[(1,0);\,(1,0);\,(0,1)]$, $\quad[y_1,y_2,y_3]=(1,0,1/2)$. The explicit solution of this problem is:

$$\theta(\varepsilon)=\mathrm{Proj}_{\|\cdot\|_2^2\le 0.5}\left[\left(\frac{1}{2+3\varepsilon},0.5\right)\right] \tag{4}$$

Removing (x_2,y_2) is equivalent to shifting ε from 0 to $-\frac{1}{3}$. IF defines the impact caused by data removal using the derivative of $\theta(\varepsilon)$:

$$IF(z_2)=\left.\frac{d\theta(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0}=-H_{\bar{\theta}}^{-1}\nabla_\theta\ell_0(x_2,y_2,\bar{\theta}),$$

where $z_2=(x_2,y_2)$, $H_{\bar{\theta}}=\frac{1}{3}\sum_{i=1}^3\nabla_{\bar{\theta}}^2\ell_0(x_i,y_i,\bar{\theta})$ is the Hessian, and $\ell_0(x_2,y_2,\bar{\theta})=(\bar{\theta}^\top x_2-y_2)^2$. IF then estimates the change $\Delta\theta$ ignoring the constraint by:

$$\theta(-\frac{1}{3})-\theta(0)\approx-\frac{1}{3}\left.\frac{d\theta(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0}.$$

However, as shown in the left panel of Figure 1, $\left.\frac{d\theta(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0}$ is not well defined because the left derivative and right derivative are inconsistent. When $\varepsilon>0$, the point $\left(\frac{1}{2+3\varepsilon},0.5\right)$ satisfies $\|\cdot\|_2^2<0.5$ and thus lies in the interior of the feasible region. When $\varepsilon<0$, $\|\theta(\varepsilon)\|_2^2$ exceeds 0.5, the solution is projected onto the ℓ_2-boundary, resulting in two different one-sided derivatives at $\varepsilon=0$. Moreover, the right panel of Figure 1 compares the ground truth $\Delta\theta$ with the change estimated by IF. When ε shifts from 0 to $-\frac{1}{3}$, θ moves along the boundary of the feasible region, whereas the estimated change by the IF significantly deviates from the ground truth due to its failure to consider the constraint. In contrast, our estimator DIF (see Definition 2), which accounts for the geometry of the feasible region, closely matches the ground truth $\Delta\theta$ along the boundary. The derivation of DIF for this toy example is provided in Appendix A.4.

3 Directional Influence Function (DIF)

To address the non-smoothness of the solution map $\theta(\varepsilon)$ w.r.t. perturbation ε and the inability of IF to handle constrained scenarios, this section proposes an alternative approach called the Directional Influence Function (DIF). We start with the definition of the classical IF.

[2]To simplify our discussion, this constrained learning formulation does not explicitly include equality constraints, which can be considered straightforwardly.

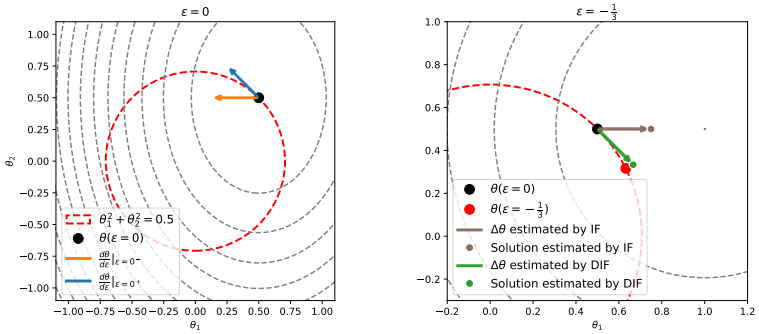


Figure 1: Loss landscape of the toy example when $\varepsilon = 0$ and $\varepsilon = -\frac{1}{3}$. The left plot illustrates the non-existence of the derivative of $\theta(\varepsilon)$ at $\varepsilon = 0$ as $\left.\frac{d\theta(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0^+} \neq \left.\frac{d\theta(\varepsilon)}{d\varepsilon}\right|_{\varepsilon=0^-}$. The right plot shows that the IF estimation of $\Delta\theta$ fails to account for the feasible region, whereas DIF accurately predicts the solution change direction and closely matches the ground truth.

Definition 1 (Influence Function)**.** The classical *IF* applies only to unconstrained problems. IF is defined as the derivative of θ w.r.t. the ε_0, i.e.,

$$IF(\bar{\theta}; Z^r) = \left.\frac{d\theta(\varepsilon_0)}{d\varepsilon_0}\right|_{\varepsilon_0=0}. \tag{5}$$

Definition 2 (Directional Influence Function)**.** *DIF* of the data subset Z^r is defined as the limiting directional derivative of the solution map $\theta(\varepsilon) : \mathbb{R}^{m+1} \to \mathbb{R}^d$ at $\bar{\varepsilon} = \mathbf{0}$, if the limit exists[3]

$$\text{DIF}\left(\bar{\theta}; Z^r\right) := D\theta(\bar{\varepsilon}; \Delta\bar{\varepsilon}) = \lim_{\substack{t \searrow 0^+ \\ \Delta\varepsilon \to \Delta\bar{\varepsilon}}} \frac{\theta(\bar{\varepsilon} + t\Delta\bar{\varepsilon}) - \theta(\bar{\varepsilon})}{t}, \tag{6}$$

where $\bar{\theta} = \theta(\bar{\varepsilon})$ denotes the model solution before perturbation. $\Delta\bar{\varepsilon} = \hat{\varepsilon} - \bar{\varepsilon} = \left[-\frac{1}{N_0}, -\frac{1}{N_1}, \ldots, -\frac{1}{N_m}\right]$ is the perturbation direction, where $\Delta\bar{\varepsilon}$ corresponds to removing the data subset Z^r from the constrained learning (1).

A function $f : X \to Y$ is called positively homogeneous if $f(\alpha x) = \alpha f(x) \quad \forall \alpha \geq 0, \forall x \in X$.

Corollary 3. *DIF is positively homogeneous,i.e.,*

$$\alpha D\theta(\bar{\varepsilon}; \Delta\bar{\varepsilon}) = D\theta(\bar{\varepsilon}; \alpha\Delta\bar{\varepsilon}), \quad \alpha \in \mathbb{R}^+$$

When ε is perturbed from $\bar{\varepsilon}$ by a finite $\Delta\bar{\varepsilon}$, IF estimates $\Delta\theta$ via a first-order Taylor expansion. IF pertains to unconstrained learning and ignores constraints; thus

$$\Delta\theta \approx \left.\frac{d\theta(\varepsilon_0)}{d\varepsilon_0}\right|_{\varepsilon=\bar{\varepsilon}} \cdot \Delta\bar{\varepsilon}_0. \tag{7}$$

This approximation, as demonstrated in the previous section, fails in the presence of constraints. Instead, DIF utilizes a directional derivative approach:

$$\Delta\theta \approx D\theta\left(\bar{\varepsilon}; \frac{\Delta\bar{\varepsilon}}{\|\Delta\bar{\varepsilon}\|}\right) \cdot \|\Delta\bar{\varepsilon}\| = D\theta\left(\bar{\varepsilon}; \Delta\bar{\varepsilon}\right). \tag{8}$$

where $D\theta(\bar{\varepsilon}; \frac{\Delta\bar{\varepsilon}}{\|\Delta\bar{\varepsilon}\|})$ is the directional derivative, and the second equation follows from Corollary 3. In what follows, we will demonstrate how to compute $D\theta\left(\bar{\varepsilon}; \Delta\bar{\varepsilon}\right)$ by solving a linearized VI system. We use $\Delta\hat{\theta}$ to denote the DIF-based estimate of $\Delta\theta$.

[3]This generalized definition accounts for potential non-smoothness of $\theta(\varepsilon)$ by allowing the direction to vary in a neighborhood of $\Delta\bar{\varepsilon}$, as is common in nonsmooth analysis[7, 37].

4 Deriving DIF via Sensitivity Analysis of Optimality Conditions

In this section, we show how DIF naturally arises from the sensitivity analysis of the optimality conditions of the constrained learning problem. Specifically, we first formulate the optimality system as a VI, then linearize this VI to obtain an auxiliary VI, and finally convert it into a QP whose solution yields the desired directional change in the model solution.

Constrained learning is typically solved using Lagrangian dual approaches, such as the augmented Lagrangian method, primal-dual methods [6, 1, 32]. To quantify the impact of data downweighting on the solution of PCL problem (2), we introduce the perturbation ε into the Lagrangian function, enabling us to analyze how the perturbation affects the optimality system. The Lagrangian function of PCL problem (2) is:

$$\begin{aligned} L(\varepsilon, \theta, \lambda) = & \frac{1}{N_0} \sum_{i=1}^{N_0} \ell_0 \left(z_i^{(0)}, \theta \right) + \varepsilon_0 \sum_{z_i^{(0)} \in Z^r} \ell_0 \left(z_i^{(0)}, \theta \right) \\ & + \sum_{j=1}^{m} \lambda_j \left[\frac{1}{N_j} \sum_{i=1}^{N_j} \ell_j \left(z_i^{(j)}, \theta \right) + \varepsilon_j \sum_{z_i^{(j)} \in Z^r} \ell_j \left(z_i^{(j)}, \theta \right) - \tau_j \right], \end{aligned} \tag{9}$$

where $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_m], \lambda_j \geq 0$ denotes the dual variables. Let $(\theta(\varepsilon), \lambda(\varepsilon))$ be the primal–dual solution for a given ε. From Section 2, we use $(\bar{\theta}, \bar{\lambda}) = (\theta(\bar{\varepsilon}), \lambda(\bar{\varepsilon}))$ to denote the optimal solution of the constrained learning problem (1).

At $\theta = \bar{\theta}$, we partition the constraints into the active set I_{Active} and the inactive set I_{Inactive}, where $I_{\text{Active}} = I_{\text{Binding}} \cup I_{\text{Non-binding}}$.

$$I_{\text{Active}} := \left\{ j \,\middle|\, \frac{1}{N_j} \sum_{i=1}^{N_j} \ell_j(z_i^{(j)}, \bar{\theta}) = \tau_j \right\}, \qquad I_{\text{Inactive}} := \left\{ j \,\middle|\, \frac{1}{N_j} \sum_{i=1}^{N_j} \ell_j(z_i^{(j)}, \bar{\theta}) < \tau_j \right\},$$

$$I_{\text{Binding}} := \{ j \in I_{\text{Active}} \mid \bar{\lambda}_j > 0 \}, \qquad I_{\text{Non-binding}} := \{ j \in I_{\text{Active}} \mid \bar{\lambda}_j = 0 \}.$$

Varying ε from $\bar{\varepsilon}$ to $\hat{\varepsilon}$ corresponds to removing Z^r. Next, we study the sensitivity of the optimality conditions with respect to ε. We begin by examining the optimality condition of PCL (2).

4.1 Optimality Condition of PCL Problem (2)

The optimality condition of PCL (2) can be described by the Karush-Kuhn-Tucker (KKT) conditions. However, directly analyzing the KKT condition of (2) turns out to be difficult for understanding how the solution changes with data perturbations. Here, we adopt a VI formulation to characterize the optimality conditions. VI is a more general form of nonlinear optimization, which allows us to analyze solution perturbations through the lens of generalized differential calculus [37]. The optimality condition of PCL (2) is:

$$-\nabla_\theta L(\varepsilon, \theta, \lambda) \in N_{\mathbb{R}^d}(\theta), \quad \nabla_\lambda L(\varepsilon, \theta, \lambda) \in N_{\mathbb{R}^m_+}(\lambda), \tag{10}$$

where $N_{\mathbb{R}^d}(\theta), N_{\mathbb{R}^m_+}(\lambda)$ denote the Normal Cone.

Definition 4 (Normal Cone;Dontchev and Rockafellar [10])**.** Let $C \subseteq \mathbb{R}^d$ be a closed convex set. The normal cone to C at a point $x \in C$ is defined as:

$$N_C(x) = \{ v \in \mathbb{R}^d : \langle v, y - x \rangle \leq 0, \ \forall y \in C \}.$$

4.2 Linear Approximation of Variational Inequality

The classical derivation of IF relies on the Taylor expansion of the first-order optimality condition, which does not naturally handle constraints. In constrained learning, a natural choice for the optimality condition is the KKT system. Yet, estimating how ε affects the KKT system is challenging due to its mixed system of equations and inequalities.

DIF adopts an alternative: representing the optimality condition as a VI (see formulation (10)) and applying a first-order approximation directly to the VI. The VI (10) can be compactly written as:

$$f(\varepsilon, \theta, \lambda) + N_E(\theta, \lambda) \ni \mathbf{0}, \tag{11}$$
$$\text{where} \quad f(\varepsilon, \theta, \lambda) = \begin{pmatrix} \nabla_\theta L(\varepsilon, \theta, \lambda) \\ -\nabla_\lambda L(\varepsilon, \theta, \lambda) \end{pmatrix}, E = \mathbb{R}^d \times \mathbb{R}^m_+$$

By definition, the solution $(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})$ satisfies the equation (11). When ε is perturbed by $\Delta\bar{\varepsilon}$, estimating $\Delta\theta$ and $\Delta\lambda$ amounts to finding $\Delta\theta$ and $\Delta\lambda$ such that

$$f(\bar{\varepsilon} + \Delta\bar{\varepsilon}, \bar{\theta} + \Delta\theta, \bar{\lambda} + \Delta\lambda) + N_E(\bar{\theta} + \Delta\theta, \bar{\lambda} + \Delta\lambda) \ni \mathbf{0} \tag{12}$$

The linear approximation of (12) is

$$f(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + \nabla_\varepsilon f(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} + \nabla_{(\theta,\lambda)} f(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \begin{bmatrix} \Delta\hat{\theta} \\ \Delta\hat{\lambda} \end{bmatrix}$$
$$+ N_E((\bar{\theta}, \bar{\lambda}) + (\Delta\hat{\theta}, \Delta\hat{\lambda})) \ni \mathbf{0}. \tag{13}$$

Here, $(\Delta\hat{\theta}, \Delta\hat{\lambda})$, which satisfies (13), serves as an approximation of $(\Delta\theta, \Delta\lambda)$. Denote

$$A := \nabla_{(\theta,\lambda)} f(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = \begin{bmatrix} \nabla^2_\theta L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}), \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \\ -\nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}), -\nabla^2_\lambda L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \end{bmatrix},$$
$$\mu := \nabla_\varepsilon f(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = \begin{bmatrix} \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \\ -\nabla_{\lambda\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \end{bmatrix},$$
$$\Delta\hat{\eta} = \begin{bmatrix} \Delta\hat{\theta} \\ \Delta\hat{\lambda} \end{bmatrix}.$$

The linearized VI (13) can be expressed in the form:

$$f(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + \mu\Delta\bar{\varepsilon} + A\Delta\hat{\eta} + N_E((\bar{\theta}, \bar{\lambda}) + \Delta\hat{\eta}) \ni 0. \tag{14}$$

Here $\Delta\bar{\varepsilon}$ is given. Note that all terms in the linearized VI (14), except for $\Delta\hat{\eta}$—which is the variable we aim to estimate—can be precomputed from the known solution $\bar{\theta}$ and $\bar{\lambda}$. The linearized VI (14) can be simplified into the following Auxiliary VI (15).

Proposition 5 (Auxiliary VI)**.** *The solution $\Delta\hat{\eta}$ to the linearized VI (14) satisfies*

$$\mu\Delta\bar{\varepsilon} + A\Delta\hat{\eta} + N_K(\Delta\hat{\eta}) \ni 0, \tag{15}$$

where $K = \mathbb{R}^d \times D$. D is a space defined as

$$D := \left\{ \Delta\hat{\lambda} \in \mathbb{R}^m \;\middle|\; \begin{array}{ll} \Delta\hat{\lambda}_j \in \mathbb{R} & \textit{for } j \in I_{\textit{Binding}}, \\ \Delta\hat{\lambda}_j \geq 0 & \textit{for } j \in I_{\textit{Non-binding}}, \\ \Delta\hat{\lambda}_j = 0 & \textit{for } j \in I_{\textit{Inactive.}} \end{array} \right\}.$$

Define (15) as the *Auxiliary VI*, which admits the same solution as the linearized VI (14), but with a simpler form.

Proposition 6. *Denote $\Delta\hat{\eta} = [\Delta\hat{\theta}; \Delta\hat{\lambda}]$ the solution to the Auxiliary VI (15). Then, $\Delta\hat{\theta}$ exactly recovers the directional derivative of the solution mapping $\theta(\varepsilon)$ at $\bar{\varepsilon}$ along $\Delta\bar{\varepsilon}$, i.e.,*

$$\Delta\hat{\theta} = D\theta(\bar{\varepsilon}; \Delta\bar{\varepsilon}). \tag{16}$$

Remark 7. By Proposition 6, computing DIF reduces to solving the Auxiliary VI (15).

The following theorem establishes the existence of DIF.

Theorem 8. *Assume the following regularity conditions hold:*

> ***Assumption 1.*** *(LICQ: Linear Independence Constraint Qualification) The gradients $\frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_\theta \ell_j(z_i^{(j)}, \bar{\theta})$ for the active constraints ($j \in I_{\textit{Activate}}$) are linearly independent.*

Assumption 2. *(SOSC: Second-Order Sufficient Condition) For any $\Delta\theta \neq 0$ such that $\Delta\theta \perp \frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_\theta \ell_j(z_i^{(j)}, \bar{\theta})$ for all $j \in I_{Binding}$, it holds that $\langle \Delta\theta, \nabla^2_{\theta\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\theta\rangle > 0$.*

Then the directional derivative $D\theta(\bar{\varepsilon}; \Delta\bar{\varepsilon})$ exists.

Proposition 9. *$\Delta\theta = \theta(\bar{\varepsilon} + \Delta\bar{\varepsilon}) - \theta(\bar{\varepsilon})$ is the ground truth and $\Delta\hat{\theta}$ is the estimation. Under the assumption of Theorem 8 , there exists a constant $M > 0$ such that, for all perturbations $\Delta\bar{\varepsilon}$ sufficiently small, the following holds:*

$$\|\Delta\theta - \Delta\hat{\theta}\| \leq M\|\Delta\bar{\varepsilon}\| \tag{17}$$

Note that LICQ in Theorem 8 is the regular constraint qualification, while SOSC essentially requires some type of "local convexity" (i.e., at the current solution) of the Lagrangian function over all feasible directions that do not violate the binding constraints. They are the conditions that the proposed DIF method can apply to analyze data attributions of a learning model. This local convexity may exist even when the model is non-convex in general, such as deep learning models. Furthermore, Proposition 9 states that, under the two assumptions, solution changes over data perturbations are Lipschitz continuous. This ensures that when constructing the auxiliary problem (18) in the next subsection, the constraint set is stable, i.e., no sudden change will occur when infinitesimally small data perturbations are imposed.

4.3 Computing DIF via Quadratic Programming

Proposition 6 implies that computing the DIF reduces to solving the Auxiliary VI (15). In this section, we construct a QP (18) whose optimality is exactly the Auxiliary VI (15). By solving this QP, we can compute DIF and estimate the directional change in solution, i.e., $\Delta\theta$. Figure 2 summarizes the derivation in this section, illustrating the connections between the PCL problem (2), VI (11), Auxiliary VI (15), and QP (18) formulations.

We provide the computational complexity analysis of the QP solver and a scalable implementation strategy in Appendix A.6.1.

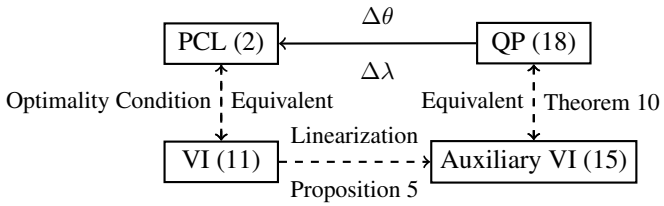


Figure 2: The relationship between (2) and (18).

We define the following QP:

$$\min_{\omega} \quad L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + \left\langle \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\bar{\varepsilon},\ \omega \right\rangle + \frac{1}{2}\left\langle \omega,\ \nabla^2_{\theta\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\,\omega \right\rangle$$

$$\text{s.t.} \quad \frac{1}{N_j}\sum_{i=1}^{N_j} \ell_j(z_i^{(j)}, \bar{\theta}) + \frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_\theta \ell_j(z_i^{(j)}, \bar{\theta})\cdot\omega + \sum_{z_i^{(j)} \in Z^r} \Delta\bar{\varepsilon}_j \ell_j(z_i^{(j)}, \bar{\theta}) - \tau_j \begin{cases} = 0, & j \in I_{\text{binding}}, \\ \leq 0, & j \in I_{\text{non-binding}}, \\ \text{free}, & j \in I_{\text{Inactive}}. \end{cases} \tag{18}$$

Theorem 10 (Auxiliary problem)**.** *Let $(\omega^\star, \zeta^\star)$ denote the optimal primal-dual solution to the QP (18). Then $(\omega^\star, \zeta^\star)$ is also a solution to the auxiliary VI (15). In particular, the QP (18) and the auxiliary VI (15) admit the same solution pair under substitution $(\omega^*, \zeta^*) = (\Delta\hat{\theta}, \Delta\hat{\lambda})$.*

Remark 11. By Proposition 6, $\Delta\hat{\theta} = D\theta(\bar{\varepsilon};\ \Delta\bar{\varepsilon})$. Hence $\omega^\star$ is identical to $\mathrm{DIF}(\bar{\theta};\ Z^r)$.

Finally, we establish the relationship between the IF and the DIF.

Proposition 12. *If no constraints are active at the solution $\bar{\theta}$, i.e., $I_{Active} = \emptyset$, then the DIF coincides with the classical IF.*

5 Validation

5.1 DIF Validating via Constrained Linear Regression

This section evaluates DIF on constrained linear regression. We generate $n = 1000$ samples with $d = 5$ features:

$$X_{ij} \sim \mathcal{N}(0,1), \qquad \theta^* \sim \mathcal{N}(0, I_d), \qquad y = X\theta^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 0.1^2 I_n).$$

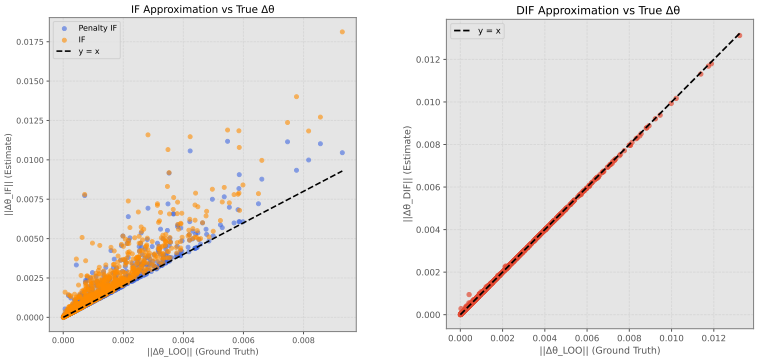
The model is trained by solving

$$\hat{\theta} = \arg\min_{\theta \in \mathbb{R}^5} \frac{1}{2n} \|X\theta - y\|_2^2 \quad \text{s.t.} \quad A_{\text{eq}}\theta = b_{\text{eq}}, \quad A_{\text{ineq}}\theta \le b_{\text{ineq}},$$

where

$$A_{\text{eq}} = [1,1,1,1,1], \quad b_{\text{eq}} = -4.0, \quad A_{\text{ineq}} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}, \quad b_{\text{ineq}} = \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}.$$

We conduct 100 single-point removal trials and compare the estimated solution change $\Delta\theta$ with leave-one-out (LOO) retraining. We compare DIF with two baselines: classical IF, which ignores constraints, and penalty-IF, which applies IF to a softly penalized surrogate. DIF is computed by solving the QP in (18); full derivations for all estimators are given in Appendix A.5. All optimization problems are solved with `CVXPY`. As shown in Figure 3, DIF nearly coincides with LOO, while IF and penalty-IF exhibit clear biases.



(a) (b)

Figure 3: Estimated versus LOO solution changes on constrained linear regression. DIF aligns closely with the $y = x$ line, whereas IF and penalty-IF show noticeable biases.

5.2 DIF Validation via Constrained CNN

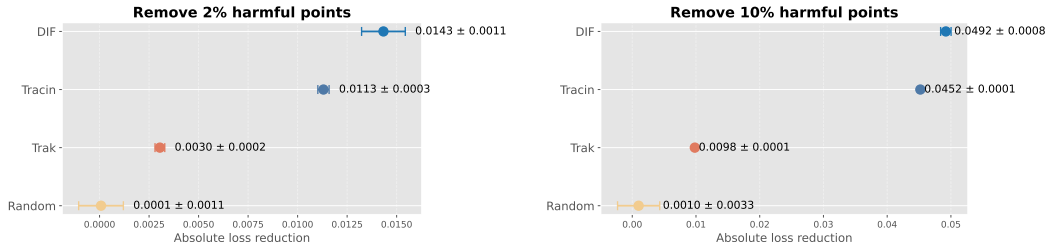
Model We adopt the following constrained learning formulation [39]:

$$\min_{\theta \in \Theta} \overline{R}(f_\theta) \quad \text{s.t.} \quad R_i(f_\theta) - \overline{R}(f_\theta) - \tau \le 0, \quad i = 1, \ldots, m, \tag{19}$$

where $R_i(f_\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_i}[\ell(f_\theta(x), y)]$ is the risk of the client (or group) i, and $\overline{R}(f_\theta) = \frac{1}{C}\sum_{i=1}^{C} R_i(f_\theta)$ is the average risk. The constraints ensure that no group's risk exceeds the global average risk by more than τ, thereby limiting the performance disparity across groups.

We use a network with seven convolutional layers and $\tanh(\cdot)$ non-linearities, modeled after the all convolutional network of Springenberg et al. [41]. We train it on the MNIST training set [26]. We solve this problem using a primal–dual method, jointly updating the model parameters θ and the Lagrange multipliers.

Heterogeneous Data Partitioning. We create non-IID group-wise data partitions following [39]. We split the dataset into C groups by allocating an α-fraction of the data uniformly at random and distributing the remaining $(1-\alpha)$-fraction in a label-skewed way, where samples are sorted by class labels and assigned consecutively to groups. Unless otherwise specified, we set the number of groups to $m = 3$ and $\alpha = 0.5$. This setup creates label-imbalanced distributions across different groups.



(a) Remove 2% harmful points

(b) Remove 10% harmful points

Figure 5: Absolute loss reduction obtained by deleting the most harmful samples selected by different methods. The dot represents the mean loss reduction, and the horizontal bar is the standard deviation.

Influence Estimation. We evaluate DIF on the top 100 influential training samples under the fairness constraint, removing one sample at a time. Since CNN retraining may move to a different local basin, we compare methods through their induced loss changes on a fixed misclassified test point rather than through parameter differences. Using $\theta' = \theta + \Delta\theta_{\mathrm{DIF}}$, we compute $\Delta\ell = R(f_{\theta'}(x_{\mathrm{test}}), y_{\mathrm{test}}) - R(f_\theta(x_{\mathrm{test}}), y_{\mathrm{test}})$. Figure 4 shows that DIF-predicted loss changes closely match leave-one-out retraining, with Pearson $r = 0.90$, suggesting that the local regularity conditions in Theorem 8 remain meaningful around the trained CNN solution.

On the same samples, standard IF fails to converge reliably due to ill-conditioned Hessian-vector products under fairness constraints. Penalty-IF converges after using a penalized surrogate, but is sensitive to the penalty coefficient and reaches only Pearson $r \approx 0.62$ after tuning. Thus, unlike DIF, penalty-IF does not accurately capture the KKT-based sensitivity structure. Additional numerical comparisons are provided in Appendix A.1.

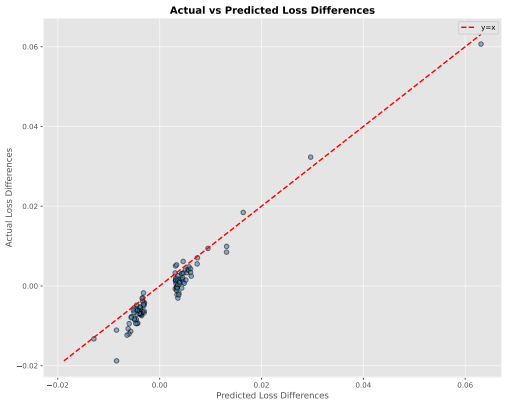


Figure 4: Actual vs. DIF-predicted loss changes on a misclassified test sample.

Counterfactual retraining. We further evaluate counterfactual retraining with DIF, constrained TracIn [34], and TRAK [33]. For each method, we compute per-sample influence scores, remove the most harmful **2%** and **10%** training samples, retrain the constrained CNN, and measure the loss change on the same misclassified test point. Constrained TracIn records Lagrangian-loss gradients along the primal-dual trajectory, while TRAK is applied to a squared-hinge penalized objective; details are provided in Appendix A.8. As shown in Figure 5, DIF and constrained TracIn substantially outperform random removal, with DIF slightly better than constrained TracIn, whereas TRAK gives little improvement. This suggests that preserving constraint-adjusted gradient information is important for identifying harmful samples in constrained training.

Additional experiments. Two additional studies are included in the appendix. Appendix A.7.1 presents a PINN-based machine unlearning case study, and Appendix A.7.2 reports a large-scale fairness-constrained ResNet-50 experiment on FairFace [22]. These results further validate DIF in downstream and large-scale constrained learning settings.

6 Related work

IF was first introduced by Hampel [17] to study the robustness of statistical estimators under distributional shifts. Decades later, Koh and Liang [23] adapted IF to machine learning to analyze the impact of individual training samples on model behavior. Since then, numerous extensions have emerged, including faster and more scalable inference methods [38, 16], and extensions to domains such as time-series forecasting [49] and generative modeling [43, 24, 28]. Beyond these, IF has been applied

to practical tasks such as defending against data poisoning [42], improving interpretability [18], and supporting machine unlearning [45, 46]. However, because IF fundamentally relies on the implicit function theorem [25], most methods require strong convexity of the model. Recent empirical studies [2] further highlight this limitation, showing that IF estimates can be fragile and often misaligned with their actual effects in practice [3, 13].

Constrained learning has gained increasing attention due to the need for machine learning models that are fair [35, 31], robust [48], and consistent with domain-specific priors [5]. For instance, invariant learning aims to ensure stable performance under distribution shifts or data transformations [4, 21], and PINNs incorporate traffic flow theory into transportation forecasting [40]. These settings highlight a growing class of learning problems where solutions must satisfy explicit feasibility requirements beyond minimizing prediction [9, 36, 19, 27, 39]. Although adding a regularization or penalty term to the objective function is common practice in ML, this approach suffers from inherent limitations. First, the choice of trade-off parameters is highly non-trivial, often requiring heuristic tuning without guarantees of correctness. Second, the gradients of the prediction loss and the penalty term rarely align; as a result, gradient-based optimization must continually balance these two gradient directions, which may lead to suboptimal updates [20, 29].

Given the wide application of constrained learning, it is essential to develop data attribution methods that explicitly account for feasibility. Classical influence-function analyses have considered constraints, but only in limited settings. [14] studied restricted minimum-divergence estimators and derived closed-form IF expressions under smooth equality constraints, while [44] extended Cook's local-influence framework to perturbations constrained by homogeneous linear restrictions. In this paper, we extend the classical IF framework to general constrained learning by analyzing solution changes within the feasible region. Specifically, we reformulate the optimality conditions as a VI and study its sensitivity with respect to data weights. This formulation naturally admits both primal and dual sensitivities, allowing us to capture how perturbations affect not only the model parameters but also the associated Lagrangian multipliers. Compared to the KKT system, the VI method offers a more general analysis framework, together with a more compact geometric language for describing the coupled dynamics of primal and dual variables by characterizing solution changes within the tangent cone of the feasible region.

7 Conclusion and Limitations

This paper introduces DIF as a constraint-aware data attribution method for constrained learning. Using VI-based sensitivity analysis, DIF derives a first-order, feasibility-preserving estimator of solution changes. Experiments on constrained linear regression and fairness-constrained CNNs show that DIF closely tracks leave-one-out retraining, whereas IF and penalty-based IF exhibit substantial bias or infeasibility. DIF may benefit applications where constraint violations are costly, such as machine unlearning, PINNs, robotics, and autonomous driving. As a limitation, DIF inherits the Hessian-related computational costs of classical IF. in large neural networks. However, existing IF acceleration techniques such as K-FAC and EK-FAC [15] can also be applied to speed up DIF.

Acknowledgements

This research was partially supported by the U.S. National Science Foundation (NSF) under Grant CMMI-2326340. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Kareem Ahmed, Tao Li, Thy Ton, Quan Guo, Kai-Wei Chang, Parisa Kordjamshidi, Vivek Srikumar, Guy Van den Broeck, and Sameer Singh. Pylon: A pytorch framework for learning with constraints. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 319–324. PMLR, 2022.

[2] Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger B Grosse. If influence functions are the answer, then what is the question? *Advances in Neural Information Processing Systems*, 35:17953–17967, 2022.

[3] Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020.

[4] Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning invariances in neural networks from training data. *Advances in neural information processing systems*, 33:17605–17616, 2020.

[5] Andrea Borghesi, Federico Baldo, and Michela Milano. Improving deep learning models via constraint-based domain knowledge: a brief survey. *arXiv preprint arXiv:2005.10691*, 2020.

[6] Luiz Chamon and Alejandro Ribeiro. Probably approximately correct constrained learning. *Advances in Neural Information Processing Systems*, 33:16722–16735, 2020.

[7] Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[8] Benjamin Coifman and Lizhe Li. A critical evaluation of the next generation simulation (ngsim) vehicle trajectory dataset. *Transportation Research Part B: Methodological*, 105:362–377, 2017.

[9] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

[10] Asen L Dontchev and R Tyrrell Rockafellar. *Implicit functions and solution mappings*, volume 543. Springer, 2009.

[11] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020.

[12] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[13] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019.

[14] Abhik Ghosh. Influence function analysis of the restricted minimum divergence estimators: A general form. *arXiv preprint arXiv:1403.6268*, 2014. Pre-print version.

[15] Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, et al. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

[16] Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. Fastif: Scalable influence functions for efficient model interpretation and debugging. *arXiv preprint arXiv:2012.15781*, 2020.

[17] Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

[18] Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. *arXiv preprint arXiv:2005.06676*, 2020.

[19] Ignacio Hounie, Alejandro Ribeiro, and Luiz FO Chamon. Resilient constrained learning. *Advances in Neural Information Processing Systems*, 36:71767–71798, 2023.

[20] Youngsik Hwang and Dongyoung Lim. Dual cone gradient descent for training physics-informed neural networks. *Advances in Neural Information Processing Systems*, 37:98563–98595, 2024.

[21] Alexander Immer, Tycho van der Ouderaa, Gunnar Rätsch, Vincent Fortuin, and Mark van der Wilk. Invariance learning in deep neural networks with differentiable laplace approximations. *Advances in Neural Information Processing Systems*, 35:12449–12463, 2022.

[22] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[23] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[24] Zhifeng Kong and Kamalika Chaudhuri. Understanding instance-based interpretability of variational auto-encoders. *Advances in Neural Information Processing Systems*, 34:2400–2412, 2021.

[25] Steven George Krantz and Harold R Parks. *The implicit function theorem: history, theory, and applications*. Springer Science & Business Media, 2002.

[26] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[27] Zihao Li, Boyi Liu, Zhuoran Yang, Zhaoran Wang, and Mengdi Wang. Double duality: Variational primal-dual policy optimization for constrained reinforcement learning. *Journal of Machine Learning Research*, 24(385):1–43, 2023.

[28] Jinxu Lin, Linwei Tao, Minjing Dong, and Chang Xu. Diffusion attribution score: Evaluating training data influence in diffusion models. *arXiv preprint arXiv:2410.18639*, 2024.

[29] Qiang Liu, Mengyu Chu, and Nils Thuerey. Config: Towards conflict-free training of physics informed neural networks. *arXiv preprint arXiv:2408.11104*, 2024.

[30] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

[31] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[32] Yatin Nandwani, Abhishek Pathak, and Parag Singla. A primal dual formulation for deep learning with constraints. *Advances in neural information processing systems*, 32, 2019.

[33] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

[34] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020.

[35] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8227–8236, 2019.

[36] Alexander Robey, Luiz Chamon, George J Pappas, Hamed Hassani, and Alejandro Ribeiro. Adversarial robustness with semi-infinite constrained learning. *Advances in Neural Information Processing Systems*, 34:6198–6215, 2021.

[37] R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

[38] Andrea Schioppa, Polina Zablotskaia, David Vilar, and Artem Sokolov. Scaling up influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8179–8186, 2022.

[39] Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro. An agnostic approach to federated learning with class imbalance. In *ICLR*, 2022.

[40] Rongye Shi, Zhaobin Mo, and Xuan Di. Physics-informed deep learning for traffic state estimation: A hybrid paradigm informed by second-order traffic models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 540–547, 2021.

[41] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[42] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, 30, 2017.

[43] Naoyuki Terashita, Hiroki Ohashi, Yuichi Nonaka, and Takashi Kanemaru. Influence estimation for generative adversarial networks. *arXiv preprint arXiv:2101.08367*, 2021.

[44] Klaus LP Vasconcellos and LM Zea Fernandez. Influence analysis with homogeneous linear restrictions. *Computational statistics & data analysis*, 53(11):3787–3794, 2009.

[45] Xin Wang, R Tyrrell Rockafellar, et al. Machine unlearning of traffic state estimation and prediction. *arXiv preprint arXiv:2507.17984*, 2025.

[46] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021.

[47] Jiangnan Xia, Yu Yang, Jiaxing Shen, Senzhang Wang, and Jiannong Cao. Fairtp: A prolonged fairness framework for traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26391–26399, 2025.

[48] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[49] Yizi Zhang, Jingyan Shen, Xiaoxue Xiong, and Yongchan Kwon. Timeinf: Time series data contribution via influence functions. *arXiv preprint arXiv:2407.15247*, 2024.

A Technical appendices and supplementary material

- A.1. Additional Penalty-IF Comparison Results
- A.2. Preliminaries: Cones and Variational Geometry
- A.3. Proofs and Auxiliary Results
  - A.3.1. Auxiliary Lemmas and Notation
  - A.3.2. Proof of Theorem 8
  - A.3.3. Proof of Proposition 5
  - A.3.4. Proof of Corollary 3
  - A.3.5. Proof of Proposition 6
  - A.3.6. Proof of Proposition 9
  - A.3.7. Proof of Proposition 10
  - A.3.8. Proof of Proposition 12
- A.4. Details in Toy Example
  - A.4.1. Classical IF: Ignoring Constraints
  - A.4.2. DIF
- A.5. Details in Constrained Linear Regression
- A.6. Computational Complexity and Scalable Implementation of the QP Solver
  - A.6.1. Computational Complexity of the DIF QP
  - A.6.2. Scalable QP Solver with Structured Curvature Approximation
- A.7. More Experiments
  - A.7.1. Machine Unlearning of Physics-Informed Neural Networks
  - A.7.2. Additional Large-Scale Validation on FairFace
- A.8. Details of Counterfactual Retraining Evaluation

A.1 Additional Penalty-IF Comparison Results

Table 1 reports representative cases from the 100 influential samples used in the constrained CNN experiment. Compared with DIF, penalty-IF exhibits larger deviations from the actual leave-one-out retraining loss changes, which is consistent with its lower Pearson correlation reported in the main text.

Table 1: Representative comparison of penalty-IF, DIF, and actual retraining loss changes on the constrained CNN experiment.

Sample ID	Predicted (Penalty-IF)	Predicted (DIF)	Actual (Retrain)
4221	0.005528	0.004979	0.003664
215	0.000016	0.004249	0.003015
1328	-0.009318	-0.003962	-0.003139
4668	0.013700	0.003922	0.004601
772	0.001171	0.003030	0.003509
4473	0.005000	0.002985	0.002487
4144	-0.007275	-0.002820	-0.000704
2828	0.002719	0.002775	0.002441
3505	0.001118	0.002747	0.003592
1163	0.001183	0.002359	0.002551

A.2 Preliminary

Definition 13 (Polar Cone; cf. [10])**.** Let $K \subseteq \mathbb{R}^n$ be a closed convex cone. The *polar cone* of K is defined as

$$K^* = \{\, y \in \mathbb{R}^n \;\mid\; \langle x, y\rangle \le 0 \;\; \forall x \in K \,\}. \tag{20}$$

Moreover, the normal vectors to K and K^* satisfy

$$y \in N_K(x) \iff x \in N_{K^*}(y) \iff x \in K,\ y \in K^*,\ \langle x, y\rangle = 0. \tag{21}$$

Definition 14 (Tangent Cone, cf. [37])**.** For a set $C \subset \mathbb{R}^n$ (not necessarily convex) and a point $x \in C$, a vector v is said to be tangent to C at x if

$$\frac{1}{\tau^k}(x^k - x) \ \to\ v \quad \text{for some } x^k \to x,\ x^k \in C,\ \tau^k \downarrow 0. \tag{22}$$

The set of all such vectors v is called the *tangent cone* to C at x and is denoted by $T_C(x)$. For $x \notin C$, we take $T_C(x) = \emptyset$.

Definition 15 (Critical Cone, cf. [37])**.** For a convex set C, any $x \in C$ and any $v \in N_C(x)$, the *critical cone* to C at x for v is

$$K_C(x, v) = \{\, w \in T_C(x) \mid w \perp v \,\}. \tag{23}$$

Definition 16 (Critical Subspaces)**.** For a convex set $C \subset \mathbb{R}^n$ and (x, v) with $v \in N_C(x)$, let $K_C(x, v)$ denote the critical cone at (x, v). Then the associated *critical subspaces* are defined as

$$K_C^+(x, v) \ =\ K_C(x, v) - K_C(x, v) \ =\ \{\, w - w' \mid w, w' \in K_C(x, v) \,\}, \tag{24}$$

$$K_C^-(x, v) \ =\ K_C(x, v) \cap [-K_C(x, v)] \ =\ \{\, w \in K_C(x, v) \mid -w \in K_C(x, v) \,\}. \tag{25}$$

Here $K_C^+(x, v)$ is the smallest linear subspace containing $K_C(x, v)$, and $K_C^-(x, v)$ is the largest linear subspace contained in $K_C(x, v)$.

Lemma 17 (Reduction Lemma;cf. 2E.4 [10])**.** *Let $C \subset \mathbb{R}^n$ be a convex set and $\bar{x} \in C$ with $\bar{v} \in N_C(\bar{x})$. Define the critical cone $K_C := K_C(\bar{x}, \bar{v})$. Then, for all sufficiently small $w, u \in \mathbb{R}^n$, we have the equivalence*

$$\bar{v} + \Delta v \in N_C(\bar{x} + \Delta x) \iff \Delta v \in N_{K_C}(\Delta x). \tag{26}$$

A.3 Proofs

A.3.1 Auxiliary Lemmas

For clarity, we first introduce some notation and rewrite the generalized equation in a standard form.

Notation. Let $\eta = (\theta, \lambda)$ denote the parameter vector, $\bar{\eta} = (\bar{\theta}, \bar{\lambda})$ be its reference value, and $\Delta\hat{\eta} = (\Delta\hat{\theta}, \Delta\hat{\lambda})$ be the estimation to $\Delta\eta = (\Delta\theta, \Delta\lambda)$.

Solution mapping. We define the solution mapping as

$$S(\varepsilon) := \{\, (\theta, \lambda) \mid f(\varepsilon, \theta, \lambda) + N_E(\theta, \lambda) \ni 0 \,\} = \{\, \eta \mid f(\varepsilon, \eta) + N_E(\eta) \ni 0 \,\}, \tag{27}$$

where $N_E(\eta)$ denotes the normal cone to E at η.

Generalized equation form. To study the local behavior of S, we rewrite the system as the generalized equation

$$G(\eta) := f(\bar{\varepsilon}, \bar{\eta}) + \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})(\eta - \bar{\eta}) + N_E(\eta), \tag{28}$$

with

$$\nabla_\eta f(\bar{\varepsilon}, \bar{\eta}) = \nabla_{(\theta,\lambda)} f(\bar{\theta}, \bar{\lambda}, \bar{x}) =: A. \tag{29}$$

At the reference point, the optimality condition (11) ensures that

$$G(\bar{\eta}) \ni 0. \tag{30}$$

Linearization. We then consider the linearized generalized equation (the first-order approximation of G):

$$G_0(\Delta\eta) := \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})\Delta\eta + N_K(\Delta\eta) = A\Delta\eta + N_K(\Delta\eta), \qquad G_0(0) \ni 0, \tag{31}$$

where $K := K_E(\bar{\eta}, -f(\bar{\varepsilon}, \bar{\eta}))$ is the critical cone. This coincides with $K = \mathbb{R}^d \times D$ as introduced in Proposition 5.

It is convenient to define the inverse-type mapping

$$\bar{s} := G_0^{-1} = (A + N_K)^{-1}, \tag{32}$$

which solves the linearized inclusion (31).

Given perturbation $u\Delta\bar{\varepsilon}$,

$$\bar{s}(-u\Delta\bar{\varepsilon}) = (A + N_K)^{-1}(-u\Delta\bar{\varepsilon}) = \{\Delta\hat{\eta} \mid A\Delta\hat{\eta} + N_K(\Delta\hat{\eta}) + u\Delta\bar{\varepsilon} \ni 0\}. \tag{33}$$

In other words, $\bar{s}(-u\Delta\bar{\varepsilon})$ yields the solution to the auxiliary VI (15).

Critical subspace of E. Following the definition 16, the critical subspaces of E at $(\bar{\theta}, \bar{\lambda}, -f(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}))$ are defined as

$$K_E^+(\bar{\theta}, \bar{\lambda}, -f(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})) = \mathbb{R}^d \times K_{\mathbb{R}^m}^+(\bar{\lambda}, \nabla_\lambda L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})), \tag{34}$$

$$K_E^-(\bar{\theta}, \bar{\lambda}, -f(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})) = \mathbb{R}^d \times K_{\mathbb{R}^m}^-(\bar{\lambda}, \nabla_\lambda L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})). \tag{35}$$

$$\Delta\hat{\eta} = (\Delta\hat{\theta}, \Delta\hat{\lambda}) \text{ satisfies } \Delta\hat{\eta} \in K_E^+ \iff \Delta\hat{\lambda}_j = 0,\ \forall j \in I_{\text{Inactive}},\ \text{and } \Delta\hat{\theta} \in \mathbb{R}^d. \tag{36}$$

Lemma 18. *Under the assumptions of Theorem 8, for any $\Delta\hat{\eta}$ satisfying the linearized VI (15), if*

$$\Delta\hat{\eta} \in K_E^+, \quad \Delta\hat{\eta} \neq 0, \quad A\Delta\hat{\eta} \perp K_E^-, \tag{37}$$

then

$$\langle \Delta\hat{\eta},\ A\Delta\hat{\eta} \rangle > 0.$$

Proof. For any $\Delta\hat{\eta}$ satisfying the linearized VI (15), it follows from the definition of the space D that $\Delta\hat{\lambda}_j = 0$ for all j in the inactive set. Therefore, any $\Delta\hat{\eta}$ satisfying VI (15) also satisfies $\Delta\hat{\eta} \in K_E^+$.

We next analyze the orthogonality condition $A\Delta\hat{\eta} \perp K_E^-$. By expanding $A(\Delta\hat{\theta}, \Delta\hat{\lambda})$ and splitting the θ- and λ-parts, we obtain:

$$\begin{aligned}
A\Delta\hat{\eta} \perp K_E^- &\iff A(\Delta\hat{\theta}, \Delta\hat{\lambda}) \perp K_E^- \\
&\iff \left(H\Delta\hat{\theta} + \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\lambda},\ -\nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta}\right) \perp K_E^- \\
&\iff H\Delta\hat{\theta} + \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\lambda} \perp \mathbb{R}^d, \quad -\nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta} \perp K_{\mathbb{R}^m}^-\left(\bar{\lambda}, \nabla_\lambda L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\right) \\
&\iff H\Delta\hat{\theta} + \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\lambda} = 0, \quad -\nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta} \perp K_{\mathbb{R}^m}^-\left(\bar{\lambda}, \nabla_\lambda L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\right).
\end{aligned}$$

Moreover, by the definition of the critical subspace, we have

$$\Delta\hat{\lambda} \in K_{\mathbb{R}^m}^-\left(\bar{\lambda}, \nabla_\lambda L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\right) \iff \Delta\hat{\lambda}_j = 0,\ \forall j \in I_{\text{non-binding}} \cup I_{\text{Inactive}}, \tag{38}$$

where I_{Inactive} and $I_{\text{non-binding}}$ denote the sets of inactive and non-binding constraints, respectively.

Consequently, the condition

$$-\nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta} \ \perp\ K_{\mathbb{R}^m}^-\left(\bar{\lambda}, \nabla_\lambda L\right)$$

only requires that

$$\nabla_{\lambda_j\theta} L\, \Delta\hat{\theta} = 0, \qquad \forall j \in I_{\text{binding}}.$$

That is,

$$A\Delta\hat{\eta} \perp K_E^- \iff H\Delta\hat{\theta} + \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\lambda} = 0, \quad \nabla_{\lambda_j\theta} L\, \Delta\hat{\theta} = 0,\ \forall j \in I_{\text{binding}}. \tag{39}$$

Notice that

$$\nabla_{\lambda_j\theta} L\, \Delta\hat{\theta} = 0 \iff \Delta\hat{\theta} \perp \frac{1}{N_j} \sum_{i=1}^{N_j} \nabla_\theta \ell_j\left(z_i^{(j)}, \bar{\theta}\right), \qquad \forall j \in I_{\text{binding}}. \tag{40}$$

Assumption 2 in Theorem 8 implies that for all $\Delta\hat{\theta}$ satisfying (39), we have

$$\left\langle \Delta\hat{\theta}, \nabla_{\theta\theta}^2 L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\theta} \right\rangle > 0. \tag{41}$$

Moreover, substituting

$$A := \nabla_{(\theta,\lambda)} f(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = \begin{bmatrix} \nabla_\theta^2 L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) & \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \\ -\nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) & -\nabla_\lambda^2 L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \end{bmatrix},$$

to (41),

we obtain

$$\langle \Delta\hat{\eta},\ A\Delta\hat{\eta}\rangle = \langle (\Delta\hat{\theta}, \Delta\hat{\lambda}),\ A(\Delta\hat{\theta}, \Delta\hat{\lambda})\rangle = \langle \Delta\hat{\theta},\ \nabla^2_{\theta\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta}\rangle. \tag{42}$$

Combining the two results yields $\langle \Delta\hat{\eta}, A\Delta\hat{\eta}\rangle > 0$.

□

Lemma 19. *Under the assumptions of Theorem 8, $\bar{s} := (A + N_K)^{-1}$ is everywhere single-valued. Equivalently, there exists a unique solution $\Delta\hat{\eta}$ to the auxiliary VI (15) given $(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})$ and $\Delta\hat{\varepsilon}$.*

Proof. We prove that $G_0^{-1} := (A + N_K)^{-1}$ is single-valued everywhere.

Assume, for contradiction, that there exist two distinct solutions $\Delta\hat{\eta}_1$ and $\Delta\hat{\eta}_2$ such that $r = G_0^{-1}(\Delta\hat{\eta}_1) = G_0^{-1}(\Delta\hat{\eta}_2)$. Then,

$$\Delta\hat{\eta}_1, \Delta\hat{\eta}_2 \in K, \qquad r - A\Delta\hat{\eta}_1 \in N_K(\Delta\hat{\eta}_1), \qquad r - A\Delta\hat{\eta}_2 \in N_K(\Delta\hat{\eta}_2).$$

This implies, by the definition of the normal cone, that

$$\langle \Delta\hat{\eta}_1,\, r - A\Delta\hat{\eta}_2\rangle \le 0, \qquad \langle \Delta\hat{\eta}_2,\, r - A\Delta\hat{\eta}_1\rangle \le 0. \tag{43}$$

Note that K is the critical cone. By condition (21), we have

$$\begin{aligned} &\Delta\hat{\eta}_1 \in K, \quad r - A\Delta\hat{\eta}_1 \in K^*, \quad \langle \Delta\hat{\eta}_1,\, r - A\Delta\hat{\eta}_1\rangle = 0, \\ &\Delta\hat{\eta}_2 \in K, \quad r - A\Delta\hat{\eta}_2 \in K^*, \quad \langle \Delta\hat{\eta}_2,\, r - A\Delta\hat{\eta}_2\rangle = 0. \end{aligned} \tag{44}$$

By the definition of the polar cone $K^* = (K^-)^\perp$, We also have

$$-A(\Delta\hat{\eta}_1 - \Delta\hat{\eta}_2) \in K^* - K^* = (K^-)^\perp, \qquad \Delta\hat{\eta}_1 - \Delta\hat{\eta}_2 \in K - K = K^+.$$

Consider the following inner product:

$$\begin{aligned} \langle \Delta\hat{\eta}_1 - \Delta\hat{\eta}_2,\ A(\Delta\hat{\eta}_1 - \Delta\hat{\eta}_2)\rangle &= \langle \Delta\hat{\eta}_1 - \Delta\hat{\eta}_2,\ [r - A\Delta\hat{\eta}_2] - [r - A\Delta\hat{\eta}_1]\rangle \\ &= \langle \Delta\hat{\eta}_1,\ r - A\Delta\hat{\eta}_2\rangle - \langle \Delta\hat{\eta}_1,\ r - A\Delta\hat{\eta}_1\rangle \\ &\quad - \langle \Delta\hat{\eta}_2,\ r - A\Delta\hat{\eta}_2\rangle + \langle \Delta\hat{\eta}_2,\ r - A\Delta\hat{\eta}_1\rangle \\ &\le 0. \end{aligned}$$

Following Lemma 18, Assumption 2 of Theorem 8 ensures that for any nonzero $\Delta\hat{\eta} \in K^+$ with $A\Delta\hat{\eta} \perp K^-$, we must have $\langle \Delta\hat{\eta},\ A\Delta\hat{\eta}\rangle > 0$. This contradicts the above inequality unless $\Delta\hat{\eta}_1 = \Delta\hat{\eta}_2$.

Therefore, the solution must be unique, which proves that $G_0^{-1} = (A + N_K)^{-1}$ is everywhere single-valued. □

Remark 20. According to Dontchev and Rockafellar [10, Chapter 2E], Consider the generalized equation $G_0 = A + N_K$, where A is a linear operator and N_K is the normal cone mapping of a closed convex cone K. If the inverse mapping G_0^{-1} is *everywhere single-valued*, then G_0^{-1} is (globally) Lipschitz continuous.

Lemma 21 (Relation between G^{-1} and G_0^{-1})**.** *Then for all $\Delta\eta$ sufficiently close to 0, the inverse mappings G^{-1} and G_0^{-1} satisfy the following relationship:*

$$G^{-1}(\gamma) = G_0^{-1}(\gamma) + \bar{\eta}, \qquad \gamma \textit{ near } 0. \tag{45}$$

Proof. Consider any $\Delta\eta$ close to 0. We first expand the generalized equation $G(\bar{\eta} + \Delta\eta)$ as

$$\begin{aligned} G(\bar{\eta} + \Delta\eta) &= f(\bar{\varepsilon}, \bar{\eta}) + \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})\, \Delta\eta + N_E(\bar{\eta} + \Delta\eta) \\ &= f(\bar{\varepsilon}, \bar{\eta}) - \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})\, \bar{\eta} + \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})(\bar{\eta} + \Delta\eta) + N_E(\bar{\eta} + \Delta\eta). \end{aligned}$$

Let $A := \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})$. Then $\gamma \in G(\bar{\eta} + \Delta\eta)$ if and only if

$$\gamma \in f(\bar{\varepsilon}, \bar{\eta}) - A\bar{\eta} + A(\bar{\eta} + \Delta\eta) + N_E(\bar{\eta} + \Delta\eta). \tag{46}$$

Rearranging the terms, this is equivalent to

$$\gamma - f(\bar{\varepsilon}, \bar{\eta}) + A\bar{\eta} \in A(\bar{\eta} + \Delta\eta) + N_E(\bar{\eta} + \Delta\eta), \tag{47}$$

which can be further rewritten as

$$\gamma - f(\bar{\varepsilon}, \bar{\eta}) \in A\Delta\eta + N_E(\bar{\eta} + \Delta\eta). \tag{48}$$

Note that $-f(\bar{\varepsilon}, \bar{\eta}) \in N_E(\bar{\eta})$. By the ***Reduction Lemma*** 17, this holds if and only if

$$\gamma \in A\Delta\eta + N_K(\Delta\eta), \tag{49}$$

which is exactly

$$\gamma \in G_0(\Delta\eta). \tag{50}$$

□

Lemma 22. *Define*

$$\sigma(\varepsilon) := G^{-1}\big(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,(\varepsilon - \bar{\varepsilon})\big) = G^{-1}\big(-u(\varepsilon - \bar{\varepsilon})\big).$$

If G^{-1} admits a single-valued Lipschitz localization around 0 for $\bar{\eta}$, then $\sigma(\varepsilon)$ is a first-order approximation of $S(\varepsilon)$ at $\bar{\varepsilon}$.

Proof. Since f is strictly differentiable at $(\bar{\varepsilon}, \bar{\eta})$, for any small perturbations $(\Delta\varepsilon, \Delta\eta)$ we have

$$f(\bar{\varepsilon} + \Delta\varepsilon, \bar{\eta} + \Delta\eta) = f(\bar{\varepsilon}, \bar{\eta}) + \nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\varepsilon + \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})\,\Delta\eta + o(\Delta\varepsilon, \Delta\eta), \tag{51}$$

where $\|o(\Delta\varepsilon, \Delta\eta)\|$ denotes the higher-order remainder term satisfying $\|o(\Delta\varepsilon, \Delta\eta)\| = o(\|\Delta\varepsilon\| + \|\Delta\eta\|)$.

By (27),

$$\begin{aligned} S(\bar{\varepsilon} + \Delta\varepsilon) &= \{\eta \mid f(\bar{\varepsilon} + \Delta\varepsilon, \eta) + N_E(\eta) \ni 0\} \\ &= \{\eta \mid f(\bar{\varepsilon}, \bar{\eta}) + \nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\Delta\varepsilon + \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})(\eta - \bar{\eta}) + o(\Delta\varepsilon, \eta - \bar{\eta}) + N_E(\eta) \ni 0\}. \end{aligned} \tag{52}$$

Then, $\eta \in S(\bar{\varepsilon} + \Delta\varepsilon)$ satisfieså that

$$f(\bar{\varepsilon}, \bar{\eta}) + \nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\varepsilon + \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})(\eta - \bar{\eta}) + o(\Delta\varepsilon, \eta - \bar{\eta}) + N_E(\eta) \ \ni\ 0. \tag{53}$$

Rearranging the terms, we obtain the equivalent inclusion

$$-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\varepsilon - o(\Delta\varepsilon, \eta - \bar{\eta}) \ \in\ f(\bar{\varepsilon}, \bar{\eta}) + \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})(\eta - \bar{\eta}) + N_E(\eta) \ =\ G(\eta), \tag{54}$$

where $G(\eta) := \nabla_\eta f(\bar{\varepsilon}, \bar{\eta})\,(\eta - \bar{\eta}) + f(\bar{\varepsilon}, \bar{\eta}) + N_E(\eta)$ denotes the linearized generalized equation in η.

Therefore, any solution $\eta \in S(\bar{\varepsilon} + \Delta\varepsilon)$ can be written as:

$$\eta \ =\ G^{-1}(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\varepsilon - o(\Delta\varepsilon, \eta - \bar{\eta}))\,, \tag{55}$$

which expresses $S(\bar{\varepsilon} + \Delta\varepsilon)$ implicitly via the inverse mapping G^{-1}.

We use κ to denote the Lipschitz constant of G^{-1}. Since $\bar{\eta} = G(0)$, we have:

we have

$$\begin{aligned} \|\eta - \bar{\eta}\| &= \|G^{-1}(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\varepsilon - o(\Delta\varepsilon, \eta - \bar{\eta})) - G^{-1}(0)\| \\ &\le k\,\big\|-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\varepsilon - o(\Delta\varepsilon, \eta - \bar{\eta})\big\|. \end{aligned} \tag{56}$$

Applying the triangle inequality gives

$$\|\eta - \bar{\eta}\| \le k\,\|\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\|\,\|\Delta\varepsilon\| + k\,\|o(\Delta\varepsilon, \eta - \bar{\eta})\|. \tag{57}$$

Since $o(\Delta\varepsilon, \eta - \bar{\eta}) = o(\|\Delta\varepsilon\| + \|\eta - \bar{\eta}\|)$ as $(\Delta\varepsilon, \eta - \bar{\eta}) \to (0, 0)$, for any $\delta > 0$ there exists a neighborhood of $(0, 0)$ such that

$$\|o(\Delta\varepsilon, \eta - \bar{\eta})\| \le \delta\big(\|\Delta\varepsilon\| + \|\eta - \bar{\eta}\|\big). \tag{58}$$

Combining (56) and (58) yields

$$\|\eta - \bar{\eta}\| \leq k \|\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\| \|\Delta\varepsilon\| + k\delta \|\Delta\varepsilon\| + k\delta \|\eta - \bar{\eta}\|. \tag{59}$$

Rearranging terms gives

$$(1 - k\delta) \|\eta - \bar{\eta}\| \leq k\big(\|\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\| + \delta\big) \|\Delta\varepsilon\|. \tag{60}$$

Finally, choosing $\delta > 0$ small enough such that $k\delta < \frac{1}{2}$, we obtain

$$\|\eta - \bar{\eta}\| \leq \frac{k}{1 - k\delta}\big(\|\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\| + \delta\big) \|\Delta\varepsilon\| = O(\|\Delta\varepsilon\|), \tag{61}$$

Since G^{-1} is Lipschitz continuous,

$$\begin{aligned} S(\bar{\varepsilon} + \Delta\varepsilon) - \sigma(\bar{\varepsilon} + \Delta\varepsilon) &= \big\|G^{-1}(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\Delta\varepsilon - o(\Delta\varepsilon, \eta - \bar{\eta})) - G^{-1}(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\Delta\varepsilon)\big\| \\ &\leq k\|o(\Delta\varepsilon, \eta - \bar{\eta})\|. \end{aligned} \tag{62}$$

Substituting (61) to (63) yields:

$$S(\bar{\varepsilon} + \Delta\varepsilon) - \sigma(\bar{\varepsilon} + \Delta\varepsilon) \leq k\|o(\Delta\varepsilon, \eta - \bar{\eta})\| = o(\|\Delta\varepsilon\|). \tag{63}$$

This proves that $\sigma(\varepsilon)$ is a first-order approximation of $S(\varepsilon)$ at $\bar{\varepsilon}$. □

A.3.2 Proof of Theorem 8

Following Lemma 19 and Remark 20, we establish that the mapping G_0^{-1} is locally Lipschitz continuous. Moreover, by Lemma 21, it holds that

$$G^{-1}(v) = G_0^{-1}(v) + \bar{\eta}$$

which implies that G^{-1} inherits the Lipschitz continuity and single-valuedness of G_0^{-1}. In this case, the Lipschitz continuity of G^{-1} ensures the applicability of Lemma 22, which shows that $\sigma(\varepsilon)$ serves as a first-order local approximation of the solution mapping $S(\varepsilon)$.

Indeed, for small $\Delta\varepsilon$,

$$\sigma(\bar{\varepsilon} + \Delta\bar{\varepsilon}) = G^{-1}\big(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\bar{\varepsilon}\big) \tag{64}$$

$$= G_0^{-1}\big(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\bar{\varepsilon}\big) + \bar{\eta}. \tag{65}$$

Therefore,

$$\begin{aligned} S(\bar{\varepsilon} + \Delta\bar{\varepsilon}) - S(\bar{\varepsilon}) &= \sigma(\bar{\varepsilon} + \Delta\bar{\varepsilon}) - \bar{\eta} + o(\|\Delta\varepsilon\|) && (66) \\ &= G_0^{-1}\big(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\bar{\varepsilon}\big) + o(\|\Delta\bar{\varepsilon}\|) \\ &= \bar{s}\big(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\bar{\varepsilon}\big) + o(\|\Delta\bar{\varepsilon}\|). && (67) \end{aligned}$$

We obtain

$$\lim_{t\downarrow 0} \frac{S(\bar{\varepsilon} + t\Delta\bar{\varepsilon}) - S(\bar{\varepsilon})}{t} = \lim_{t\downarrow 0} \frac{\bar{s}\big(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,t\Delta\bar{\varepsilon}\big) + o(\|\Delta\bar{\varepsilon}\|)}{t} \tag{68}$$

$$\tag{69}$$

Since N_K is the normal cone mapping of the convex cone K, it satisfies $N_K(\alpha w) = \alpha N_K(w)$ for all $\alpha > 0$. Together with the linearity of A, this gives $(A + N_K)(\alpha w) = \alpha (A + N_K)(w)$. Hence $\bar{s} = (A + N_K)^{-1}$ is positively homogeneous,

$$\lim_{t\downarrow 0} \frac{\bar{s}\big(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,t\Delta\bar{\varepsilon}\big) + o(\|\Delta\bar{\varepsilon}\|)}{t} = \bar{s}(-\nabla_\varepsilon f(\bar{\varepsilon}, \bar{\eta})\,\Delta\bar{\varepsilon}). \tag{70}$$

Hence, S is directionally differentiable at $\bar{\varepsilon}$ with

$$DS(\bar{\varepsilon})(\Delta\bar{\varepsilon}) = \bar{s}\big(-\nabla_{\varepsilon} f(\bar{\varepsilon}, \bar{\eta})\,\Delta\bar{\varepsilon}\big).$$

Equation (33) ensures that

$$\bar{s}\left(-\nabla_{\varepsilon} f(\bar{\varepsilon}, \bar{\eta})\Delta\bar{\varepsilon}\right)$$

is the solution of the auxiliary VI (15). This implies that, under the assumptions of Theorem 8, the solution mapping S is directionally differentiable at $\bar{\varepsilon}$, and its directional derivative is given by the solution of the auxiliary VI (15). Since $S(\varepsilon) = (\theta, \lambda)$, the DIF $D\theta(\bar{\varepsilon}; \Delta\bar{\varepsilon})$ defined in Definition 2 corresponds to the θ-component of$DS(\bar{\varepsilon})(\Delta\bar{\varepsilon})$. This guarantees the existence of the DIF.

A.3.3 Proof of Proposition 5

Proof. Given $(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})$ satisfy the optimality condition (11), we have

$$\nabla_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + N_{\mathbb{R}^d}(\bar{\theta}) \ni 0, \quad -\nabla_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + N_{\mathbb{R}^m_+}(\bar{\lambda}) \ni 0. \tag{71}$$

Following the definition of the normal cone (Def. 4), condition (71) implies that:

$$\theta \in \mathbb{R}^d, \nabla_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = 0, \tag{72}$$

$$\lambda \in \mathbb{R}^m_+, -\nabla_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \leq 0. \tag{73}$$

Note that R^m_+ is a closed, convex cone. Following condition (21), we have

$$\nabla_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \cdot \bar{\lambda} = 0, \tag{74}$$

which is consistent with the complementary slackness condition in the KKT system.

Now we derive the equivalent of VI (13) and VI (15). We assume the $\Delta\bar{\varepsilon}, \Delta\hat{\theta}, \Delta\hat{\lambda}$ are sufficiently close to 0.

VI (13) requires that

$$\nabla_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\bar{\varepsilon} + \nabla^2_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\theta} + \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\lambda} + N_{\mathbb{R}^d}(\bar{\theta} + \Delta\hat{\theta}) \ni 0$$

$$\Updownarrow$$

$$\bar{\theta} + \Delta\hat{\theta} \in \mathbb{R}^d, \quad \nabla_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\bar{\varepsilon} + \nabla^2_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\theta} + \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\lambda} = 0 \tag{75}$$

and

$$-\nabla_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) - \nabla_{\lambda\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\bar{\varepsilon} - \nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\theta} - \nabla^2_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\lambda} + N_{\mathbb{R}^m_+}(\bar{\lambda} + \Delta\hat{\lambda}) \ni 0$$

$$\Updownarrow$$

$$\bar{\lambda} + \Delta\hat{\lambda} \in \mathbb{R}^m_+, \quad \nabla_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + \nabla_{\lambda\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\bar{\varepsilon} + \nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\theta} + \nabla^2_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\lambda} \leq 0 \tag{76}$$

It is straightforward to verify that $\nabla^2_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = 0$. By equation (72), $\nabla_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = 0$. Substituting $\nabla_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = 0$ to (75) and $\nabla^2_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = 0$ to (76) yields

$$\bar{\theta} + \Delta\hat{\theta} \in \mathbb{R}^d, \quad \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\bar{\varepsilon} + \nabla^2_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\theta} + \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\lambda} = 0, \tag{77}$$

and

$$\bar{\lambda} + \Delta\hat{\lambda} \in \mathbb{R}^m_+, \quad \nabla_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + \nabla_{\lambda\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\bar{\varepsilon} + \nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\Delta\hat{\theta} \leq 0 \tag{78}$$

Since $\mathbb{R}^m_+$ is a closed convex cone, by (21), we have

$$\big(\nabla_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + \nabla_{\lambda\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\,\Delta\bar{\varepsilon} + \nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\,\Delta\hat{\theta}\big) \cdot (\bar{\lambda} + \Delta\hat{\lambda}) = 0 \tag{79}$$

In summary, VI (13) is equivalent to the system consisting of (77), (78), and (79). Note that $\lambda = [\lambda_1, \ldots, \lambda_j]$, we now further discuss the formulation (78) and (79) for $j \in I_{\text{Inactive}}$, $j \in I_{\text{Binding}}$, and $j \in I_{\text{Non-binding}}$.

Case 1. If $j \in I_{\text{Inactive}}$, then $\nabla_{\lambda_j} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) < 0$ and $\bar{\lambda}_j = 0$. Since $\Delta\bar{\varepsilon}$, $\Delta\hat{\theta}$, and $\Delta\hat{\lambda}$ are all sufficiently close to 0 and $\nabla_{\lambda_j} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) < 0$, condition (78)

$$\nabla_{\lambda_j} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) + \nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\bar{\varepsilon} + \nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\theta} \leq 0$$

is automatically satisfied. Therefore, the term

$$\nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} + \nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta} \ \text{ is free.} \tag{80}$$

Moreover, condition (79) implies that

$$\Delta\hat{\lambda}_j = 0 \quad \text{for } j \in I_{\text{Inactive}}. \tag{81}$$

Case 2. If $j \in I_{\text{Non-binding}}$, then $\nabla_{\lambda_j} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = 0$ and $\bar{\lambda}_j = 0$. Substituting $\nabla_{\lambda_j} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = 0$ and $\bar{\lambda}_j = 0$ into condition (78) yields

$$\begin{aligned} \Delta\hat{\lambda}_j &\geq 0 \qquad && \text{for } j \in I_{\text{Non-binding}}, \\ \nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} + \nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta} &\leq 0 \qquad && \text{for } j \in I_{\text{Non-binding}}. \end{aligned} \tag{82}$$

Case 3. If $j \in I_{\text{Binding}}$, then $\nabla_{\lambda_j} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) = 0$ and $\bar{\lambda}_j \geq 0$. Since $\Delta\hat{\lambda}$ is close to 0, $\bar{\lambda} + \Delta\hat{\lambda} \in \mathbb{R}^m_+$ is automatically satisfied. To satisfy conditions (78) and (79), we must have

$$\nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} + \nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta} = 0.$$

In summary, we obtain

$$\begin{aligned} \Delta\hat{\lambda}_j \ \text{ is free.} \qquad && \text{for } j \in I_{\text{Binding}}, \\ \nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} + \nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta} = 0 \qquad && \text{for } j \in I_{\text{Binding}}. \end{aligned} \tag{83}$$

Taken together, Cases 1–3 show that VI (13) is equivalent to the system consisting of (77) and (80–83).On the other hand, the following argument shows that VI (15) is also equivalent to this system.

The VI (15) implies that:

$$\begin{aligned} & \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\bar{\varepsilon} + \nabla^2_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\theta} + \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\lambda} + N_{\mathbb{R}^d}(\Delta\hat{\theta}) \ni 0 \\ & \Updownarrow \\ & \Delta\hat{\theta} \in \mathbb{R}^d, \quad \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\bar{\varepsilon} + \nabla^2_{\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\theta} + \nabla_{\theta\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\lambda} = 0 \end{aligned} \tag{84}$$

and

$$\begin{aligned} & -\nabla_{\lambda\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\bar{\varepsilon} - \nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\theta} - \nabla^2_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\lambda} + N_D(\Delta\hat{\lambda}) \ni 0 \\ & \Updownarrow \\ & \Delta\hat{\lambda} \in D, \quad (-\nabla_{\lambda\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\bar{\varepsilon} - \nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\theta} - \nabla^2_{\lambda} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\lambda})(\Delta\hat{\lambda}' - \Delta\hat{\lambda}) \leq 0, \forall \Delta\hat{\lambda}' \in D \end{aligned} \tag{85}$$

Since D is a closed convex cone, by (21), we have

$$\left(\nabla_{\lambda\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} + \nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta}\right) \cdot \Delta\hat{\lambda} = 0 \tag{86}$$

Case 1. If $j \in I_{\text{Inactive}}$, then by the definition of the space D (see Proposition 5), we have

$$\Delta\hat{\lambda}_j = 0 \ \ , \forall\, \Delta\hat{\lambda} \in D \tag{87}$$

Thus, condition (85) is automatically satisfied. Therefore, the term

$$\nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} + \nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\hat{\theta} \text{ is free.} \tag{88}$$

Case 2. If $j \in I_{\text{Non-binding}}$, then by the definition of the space D (see Proposition 5), we have

$$\Delta\hat{\lambda}_j \geq 0, \quad \forall\, \Delta\hat{\lambda} \in D \tag{89}$$

By the condition (86),

$$\nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\bar{\varepsilon} + \nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\theta} \leq 0 \quad \text{for } j \in I_{\text{Non-binding}}\,. \tag{90}$$

Case 3. If $j \in I_{\text{binding}}$, then by the definition of the space D (see Proposition 5), we have

$$\Delta\hat{\lambda}_j \quad \text{is free} \tag{91}$$

By the condition (86),

$$\nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\bar{\varepsilon} + \nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}) \Delta\hat{\theta} = 0 \quad \text{for } j \in I_{\text{Binding}} \tag{92}$$

Since (84,87–92) and (77,80–83) are equivalent formulations, VI (15) is equivalent to VI (13). □
□

A.3.4 Proof of corollary 3

Proof. Recall the directional derivative

$$D\theta(\bar{\varepsilon}; v) := \lim_{t \downarrow 0, v' \to v} \frac{\theta\left(\bar{\varepsilon} + t v'\right) - \theta(\bar{\varepsilon})}{t},$$

whenever the limit exists and is finite.
Let $\alpha \geq 0$. - If $\alpha = 0$: by the definition with $v \equiv 0$,

$$D\theta(\bar{\varepsilon}; 0) = \lim_{t \downarrow 0, v' \to 0} \frac{\theta\left(\bar{\varepsilon} + t v'\right) - \theta(\bar{\varepsilon})}{t} = 0,$$

hence $0 \cdot D\theta(\bar{\varepsilon}; v) = D\theta(\bar{\varepsilon}; 0)$. - If $\alpha > 0$: using the definition with the direction αv,

$$D\theta(\bar{\varepsilon}; \alpha v) = \lim_{t \downarrow 0, u' \to \alpha v} \frac{\theta\left(\bar{\varepsilon} + t u'\right) - \theta(\bar{\varepsilon})}{t}.$$

Choose $u' = \alpha v'$ with $v' \to v$. Then

$$\begin{aligned} D\theta(\bar{\varepsilon}; \alpha v) &= \lim_{t \downarrow 0, v' \to v} \frac{\theta\left(\bar{\varepsilon} + t\left(\alpha v'\right)\right) - \theta(\bar{\varepsilon})}{t} = \lim_{t \downarrow 0, v' \to v} \frac{\theta\left(\bar{\varepsilon} + (\alpha t) v'\right) - \theta(\bar{\varepsilon})}{t} \\ &= \lim_{s \downarrow 0, v' \to v} \frac{\theta\left(\bar{\varepsilon} + s v'\right) - \theta(\bar{\varepsilon})}{s/\alpha} \quad (\text{ set } s = \alpha t) \\ &= \alpha \lim_{s \downarrow 0, v' \to v} \frac{\theta\left(\bar{\varepsilon} + s v'\right) - \theta(\bar{\varepsilon})}{s} = \alpha D\theta(\bar{\varepsilon}; v). \end{aligned}$$

Combining the two cases yields

$$\alpha D\theta(\bar{\varepsilon}; \Delta\bar{\varepsilon}) = D\theta(\bar{\varepsilon}; \alpha\Delta\bar{\varepsilon}), \quad \forall \alpha \in \mathbb{R}^+$$

□

A.3.5 Proof of proposition 6

Section A.3.2 has proved that the directional derivative of the solution mapping S is characterized by the auxiliary VI (15). Specifically,

$$DS(\bar{\varepsilon})(\Delta\bar{\varepsilon}) = \Delta\hat{\eta} \quad \text{where} \quad \mu\,\Delta\bar{\varepsilon} + A\,\Delta\hat{\eta} + N_K(\Delta\hat{\eta}) \ni 0. \tag{93}$$

Since the DIF $D\theta(\bar{\varepsilon};\Delta\bar{\varepsilon})$ defined in Definition 2 is the θ-component of $DS(\bar{\varepsilon})(\Delta\bar{\varepsilon})$, we have

$$\Delta\hat{\theta} \;=\; D\theta(\bar{\varepsilon};\Delta\bar{\varepsilon}). \tag{94}$$

A.3.6 Proof of Proposition 9

Proof. Let $\Delta\hat{\eta} = [\Delta\hat{\theta}, \Delta\hat{\lambda}]$ denote the solution of Auxiliary VI (15). By (33),

$$\Delta\hat{\eta} \;=\; \bar{s}(-u\,\Delta\bar{\varepsilon})\,.$$

By definition,

$$\Delta\eta \;=\; \eta - \bar{\eta} \;=\; S(\bar{\varepsilon}+\Delta\bar{\varepsilon}) - \bar{\eta}.$$

Therefore,

$$\|\Delta\theta - \Delta\hat{\theta}\| \;\le\; \|\Delta\eta - \Delta\hat{\eta}\| \;=\; \big\|S(\bar{\varepsilon}+\Delta\bar{\varepsilon}) - \bar{\eta} - \bar{s}(-u\,\Delta\bar{\varepsilon})\big\|. \tag{95}$$

By Lemma 22,

$$\sigma(\varepsilon) := G^{-1}\Big(-\nabla_\varepsilon f(\bar{\varepsilon},\bar{\eta})\,(\varepsilon-\bar{\varepsilon})\Big)$$

is a first-order approximation of $S(\varepsilon)$ and $\sigma(\bar{\varepsilon}) = S(\bar{\varepsilon}) = \bar{\eta}$. Thus,

$$S(\bar{\varepsilon}+\Delta\bar{\varepsilon}) = \sigma(\bar{\varepsilon}+\Delta\bar{\varepsilon}) + o(\|\Delta\bar{\varepsilon}\|). \tag{96}$$

By Lemma 21,

$$\sigma(\bar{\varepsilon}+\Delta\bar{\varepsilon}) = G^{-1}(-u\,\Delta\bar{\varepsilon}) = G_0^{-1}(-u\,\Delta\bar{\varepsilon}) + \bar{\eta} = \bar{s}(-u\,\Delta\bar{\varepsilon}) + \bar{\eta}. \tag{97}$$

Substituting (97) and (96) into right side of (95) yields

$$\big\|S(\bar{\varepsilon}+\Delta\bar{\varepsilon}) - \bar{\eta} - \bar{s}(-u\,\Delta\bar{\varepsilon})\big\| = o(\|\Delta\bar{\varepsilon}\|),$$

hence

$$\|\Delta\theta - \Delta\hat{\theta}\| \le o(\|\Delta\bar{\varepsilon}\|).$$

By the little-o definition, for every $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that, whenever $0 < \|\Delta\bar{\varepsilon}\| < \delta(\epsilon)$,

$$\frac{\|\Delta\theta - \Delta\hat{\theta}\|}{\|\Delta\bar{\varepsilon}\|} < \epsilon. \tag{98}$$

Fix any $M > 0$ and set $\epsilon := M$. Then there exists $\delta := \delta(M) > 0$ such that, for all $0 < \|\Delta\bar{\varepsilon}\| < \delta$,

$$\|\Delta\theta - \Delta\hat{\theta}\| \le M\,\|\Delta\bar{\varepsilon}\|. \tag{99}$$

□

A.3.7 Proof of Theorem 10

Define the Lagrangian of QP (18) as

$$\begin{aligned}
\mathcal{L}_{QP}(\omega,\zeta) = L(\bar{\varepsilon},\bar{\theta},\bar{\lambda}) &+ \big\langle \nabla_{\theta\varepsilon} L(\bar{\varepsilon},\bar{\theta},\bar{\lambda})\cdot\Delta\bar{\varepsilon},\ \omega\big\rangle + \frac{1}{2}\big\langle \omega,\ \nabla^2_{\theta\theta} L(\bar{\varepsilon},\bar{\theta},\bar{\lambda})\,\omega\big\rangle \\
&+ \sum_{j\in I_{\text{binding}}} \zeta_j \left[\frac{1}{N_j}\sum_{i=1}^{N_j} \ell_j(z_i^{(j)},\bar{\theta}) + \frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_\theta \ell_j(z_i^{(j)},\bar{\theta})\cdot\omega + \sum_{z_i^{(j)}\in Z^r} \Delta\bar{\varepsilon}_j\,\ell_j(z_i^{(j)},\bar{\theta}) - \tau_j\right] \\
&+ \sum_{j\in I_{\text{non-binding}}} \zeta_j \left[\frac{1}{N_j}\sum_{i=1}^{N_j} \ell_j(z_i^{(j)},\bar{\theta}) + \frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_\theta \ell_j(z_i^{(j)},\bar{\theta})\cdot\omega + \sum_{z_i^{(j)}\in Z^r} \Delta\bar{\varepsilon}_j\,\ell_j(z_i^{(j)},\bar{\theta}) - \tau_j\right] \\
&+ \sum_{j\in I_{\text{Inactive}}} \zeta_j \left[\frac{1}{N_j}\sum_{i=1}^{N_j} \ell_j(z_i^{(j)},\bar{\theta}) + \frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_\theta \ell_j(z_i^{(j)},\bar{\theta})\cdot\omega + \sum_{z_i^{(j)}\in Z^r} \Delta\bar{\varepsilon}_j\,\ell_j(z_i^{(j)},\bar{\theta}) - \tau_j\right].
\end{aligned} \tag{100}$$

A pair $(\omega^\star, \zeta^\star)$ satisfies the Karush–Kuhn–Tucker (KKT) conditions if:

(i) Primal feasibility.

$$\frac{1}{N_j}\sum_{i=1}^{N_j}\ell_j(z_i^{(j)},\bar{\theta}) + \frac{1}{N_j}\sum_{i=1}^{N_j}\nabla_\theta\ell_j(z_i^{(j)},\bar{\theta})\cdot\omega^\star + \sum_{z_i^{(j)}\in Z^r}\Delta\bar{\varepsilon}_j\,\ell_j(z_i^{(j)},\bar{\theta}) - \tau_j = 0, \qquad \forall j \in I_{\text{binding}}, \tag{101}$$

$$\frac{1}{N_j}\sum_{i=1}^{N_j}\ell_j(z_i^{(j)},\bar{\theta}) + \frac{1}{N_j}\sum_{i=1}^{N_j}\nabla_\theta\ell_j(z_i^{(j)},\bar{\theta})\cdot\omega^\star + \sum_{z_i^{(j)}\in Z^r}\Delta\bar{\varepsilon}_j\,\ell_j(z_i^{(j)},\bar{\theta}) - \tau_j \le 0, \qquad \forall j \in I_{\text{non-binding}}. \tag{102}$$

For $j \in I_{\text{Inactive}}$, there is no constraint.

(ii) Stationarity.

$$\nabla_{\theta\varepsilon}L(\bar{\varepsilon},\bar{\theta},\bar{\lambda})\cdot\Delta\bar{\varepsilon} + \nabla^2_{\theta\theta}L(\bar{\varepsilon},\bar{\theta},\bar{\lambda})\,\omega^\star + \sum_{j\in I_{\text{binding}}\cup I_{\text{non-binding}}\cup I_{\text{Inactive}}}\zeta_j^\star\left(\frac{1}{N_j}\sum_{i=1}^{N_j}\nabla_\theta\ell_j(z_i^{(j)},\bar{\theta})\right) = 0. \tag{103}$$

(iii) Dual feasibility.

$$\zeta_j^\star \ge 0, \qquad \forall j \in I_{\text{non-binding}}. \tag{104}$$

(iv) Complementary slackness.

$$\zeta_j^\star\left[\frac{1}{N_j}\sum_{i=1}^{N_j}\ell_j(z_i^{(j)},\bar{\theta}) + \frac{1}{N_j}\sum_{i=1}^{N_j}\nabla_\theta\ell_j(z_i^{(j)},\bar{\theta})\cdot\omega^\star + \sum_{z_i^{(j)}\in Z^r}\Delta\bar{\varepsilon}_j\,\ell_j(z_i^{(j)},\bar{\theta}) - \tau_j\right] = 0, \qquad \forall j \in I_{\text{non-binding}}. \tag{105}$$

We now show that any $(\omega^\star, \zeta^\star)$ satisfying the above KKT conditions also satisfies the auxiliary VI (15). Recall that in Section A.3.3, we have proved that the auxiliary VI (15) is equivalent to the system $(84, 87--92)$.

Note that

$$\begin{aligned} L(\varepsilon,\theta,\lambda) = {} & \frac{1}{N_0}\sum_{i=1}^{N_0}\ell_0(z_i^{(0)},\theta) + \varepsilon_0\sum_{z_i^{(0)}\in Z^r}\ell_0(z_i^{(0)},\theta) \\ & + \sum_{j=1}^{m}\lambda_j\left[\frac{1}{N_j}\sum_{i=1}^{N_j}\ell_j(z_i^{(j)},\theta) + \varepsilon_j\sum_{z_i^{(j)}\in Z^r}\ell_j(z_i^{(j)},\theta) - \tau_j\right], \end{aligned}$$

we have

$$\frac{1}{N_j}\sum_{i=1}^{N_j}\nabla_\theta\ell_j(z_i^{(j)},\bar{\theta}) = \nabla_{\theta\lambda_j}L(\bar{\varepsilon},\bar{\theta},\bar{\lambda}). \tag{106}$$

The stationarity condition can be rewritten as

$$\nabla_{\theta\varepsilon}L(\bar{\varepsilon},\bar{\theta},\bar{\lambda})\,\Delta\bar{\varepsilon} + \nabla^2_\theta L(\bar{\varepsilon},\bar{\theta},\bar{\lambda})\,\omega^\star + \nabla_{\theta\lambda}L(\bar{\varepsilon},\bar{\theta},\bar{\lambda})\,\zeta^\star = 0. \tag{107}$$

By the definition of the normal cone (Definition 4), this is equivalent to

$$\nabla_{\theta\varepsilon}L(\bar{\varepsilon},\bar{\theta},\bar{\lambda})\,\Delta\bar{\varepsilon} + \nabla^2_\theta L(\bar{\varepsilon},\bar{\theta},\bar{\lambda})\,\omega^\star + \nabla_{\theta\lambda}L(\bar{\varepsilon},\bar{\theta},\bar{\lambda})\,\zeta^\star \in N_{\mathbb{R}^d}(\Delta\hat{\theta}). \tag{108}$$

We distinguish the three index sets.

Case 1 ($j \in I_{\text{Inactive}}$). By complementary slackness, $\zeta_j^\star = 0$.

Case 2 ($j \in I_{\text{non-binding}}$). By complementary slackness, $\zeta_j^\star \geq 0$. Moreover, since $\frac{1}{N_j}\sum_{i=1}^{N_j} \ell_j(z_i^{(j)}, \bar{\theta}) = \tau_j$, substituting $\nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} = \sum_{z_i^{(j)} \in Z^r} \Delta\bar{\varepsilon}_j \ell_j(z_i^{(j)}, \bar{\theta})$ and $\nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \omega^\star = \frac{1}{N_j}\sum_{i=1}^{N_j} \nabla_\theta \ell_j(z_i^{(j)}, \bar{\theta})\omega^\star$ into the primal feasibility condition gives

$$\nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} + \nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \omega^* \leq 0. \tag{109}$$

Case 3 ($j \in I_{\text{binding}}$). Here $\zeta_j^\star$ is free. Since $\frac{1}{N_j}\sum_{i=1}^{N_j} \ell_j(z_i^{(j)}, \bar{\theta}) = \tau_j$, using the same substitutions into of case 3 into primal feasibility as above yields

$$\nabla_{\lambda_j \varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} + \nabla_{\lambda_j \theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \omega^* = 0. \tag{110}$$

Combining all cases, $\zeta^\star$ satisfies

$$\nabla_{\lambda\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon} - \nabla_{\lambda\theta} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \omega^\star - \nabla_\lambda^2 L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \zeta^\star + N_{D'}(\zeta^\star) \ni 0, \tag{111}$$

where

$$D' := \Big\{ \zeta^* \in \mathbb{R}^m \;\Big|\; \zeta_j^* \geq 0 \;(j \in I_{\text{non-binding}}),\; \zeta_j^* = 0 \;(j \in I_{\text{Inactive}}) \Big\}.$$

(111) together with (107) shows that $(\omega^\star, \zeta^\star)$ satisfies

$$0 \in A(\omega^\star, \zeta^\star) + \mu\Delta\bar{\varepsilon} + N_K(\omega^\star, \zeta^\star),$$

which is exactly the auxiliary VI (15). □

A.3.8 Proof of Proposition 12

Proof. By Theorem 10 we have $\omega^\star = \Delta\hat{\theta}$, where $\Delta\hat{\theta}$ is the θ-component of the solution to the auxiliary VI (15). By Proposition 6, $\Delta\hat{\theta} = D\theta(\bar{\varepsilon}; \Delta\bar{\varepsilon})$. Therefore, ω^* is the DIF.

Assume that no constraints are active at $\bar{\theta}$. Then the KKT conditions reduce to the unconstrained optimality system, and the QP (18) has the unique solution $\omega^\star$ given by the Newton/IF direction:

$$\omega^\star \;=\; -\big[\nabla_{\theta\theta}^2 L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\big]^{-1} \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda}). \tag{112}$$

Since "no active constraints" implies $\bar{\lambda} = 0$, substituting $\bar{\lambda} = 0$ into (112) yields

$$\omega^\star = -\big[\nabla_{\theta\theta}^2 L(\bar{\varepsilon}, \bar{\theta}, 0)\big]^{-1} \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, 0) \;=\; \frac{d\,\theta(\varepsilon)}{d\varepsilon_0}\Big|_{\varepsilon_0 = 0},$$

which is exactly the *influence function* (IF) defined in Koh and Liang [23].

Therefore, in the inactive-constraint case, the QP solution $\omega^\star$ coincides with the IF, as claimed. □

A.4 Details in Toy example

We consider the toy example (3)

$$\min_{\theta \in \mathbb{R}^2} \; \frac{1}{3}(\theta^\top x_1 - y_1)^2 + \frac{1}{3}(\theta^\top x_2 - y_2)^2 + \frac{1}{3}(\theta^\top x_3 - y_3)^2 + \varepsilon\,(\theta^\top x_2 - y_2)^2 \quad \text{s.t.} \quad \|\theta\|_2^2 \leq 1,$$

with data $x_1 = (1, 0)$, $x_2 = (1, 0)$, $x_3 = (0, 1)$ and $y_1 = 1$, $y_2 = 0$, $y_3 = \frac{1}{2}$. At $\varepsilon = 0$, the constrained solution is $\bar{\theta} = (0.5,\, 0.5)$. The corresponding dual variable is $\bar{\lambda} = 0$.

Gradients and Hessian. For a single squared loss $\ell_i(\theta) = (\theta^\top x_i - y_i)^2$,

$$\nabla_\theta \ell_i(\theta) = 2(\theta^\top x_i - y_i)x_i, \qquad \nabla_{\theta\theta}^2 \ell_i(\theta) = 2x_i x_i^\top.$$

Hence, at $\varepsilon = 0$ the (unconstrained) Hessian of $\frac{1}{3}\sum_{i=1}^3 \ell_i$ is

$$H = \frac{1}{3}\sum_{i=1}^{3} 2x_i x_i^\top = \frac{2}{3}\big(x_1 x_1^\top + x_2 x_2^\top + x_3 x_3^\top\big) = \begin{pmatrix} \frac{4}{3} & 0 \\ 0 & \frac{2}{3} \end{pmatrix}. \tag{113}$$

The mixed derivative of the perturbed part is

$$\nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}) = \nabla_\theta \ell_2(\bar{\theta}) = 2(\bar{\theta}^\top x_2 - y_2)x_2 = 2 \cdot 0.5 \cdot (1, 0) = (1, 0). \tag{114}$$

A.4.1 Classical IF (ignoring constraints)

Differentiating the stationarity condition $\nabla_\theta L(\varepsilon, \theta) = 0$ w.r.t. ε at $(\bar{\varepsilon}, \bar{\theta})$ gives

$$H \frac{d\theta}{d\varepsilon} + \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}) = 0 \quad \Rightarrow \quad \frac{d\theta}{d\varepsilon} = -H^{-1} \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}).$$

Using equation 113–equation 114,

$$\frac{d\theta}{d\varepsilon} = -\begin{pmatrix} \frac{3}{4} & 0 \\ 0 & \frac{3}{2} \end{pmatrix} (1, 0) = \Big(-\tfrac{3}{4},\, 0 \Big).$$

Removing sample z_2 corresponds to $\Delta\varepsilon = -\frac{1}{3}$, so the IF estimate is

$$\Delta\theta_{\mathrm{IF}} \approx \frac{d\theta}{d\varepsilon}\, \Delta\varepsilon = \Big(-\tfrac{3}{4},\, 0 \Big) \cdot \Big(-\tfrac{1}{3} \Big) = \Big(\tfrac{1}{4},\, 0 \Big). \tag{115}$$

Then $\bar{\theta} + \Delta\theta_{\mathrm{IF}} = (0.75,\, 0.5)$ whose ℓ_1-norm equals 1.25, i.e., the IF step is infeasible.

A.4.2 DIF (feasible, sensitivity on the linearized VI)

DIF linearizes the KKT/VI system at $(\bar{\varepsilon}, \bar{\theta})$ and searches $\Delta\theta$ in the tangent subspace. The QP (18) corresponding to the toy problem (3) is

$$\min_{\Delta\theta \in \mathbb{R}^2} \ \big\langle \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta})\, \Delta\varepsilon,\ \Delta\theta \big\rangle + \frac{1}{2}\, \Delta\theta^\top H\, \Delta\theta \quad \text{s.t.} \quad \Delta\theta_1 + \Delta\theta_2 = 0. \tag{116}$$

With $\Delta\varepsilon = -\frac{1}{3}$ and $b := \nabla_{\theta\varepsilon} L\, \Delta\varepsilon = (-\frac{1}{3}, 0)$, (116) is

$$\min_{\Delta\theta}\ b^\top \Delta\theta + \tfrac{1}{2}\, \Delta\theta^\top H\, \Delta\theta \quad \text{s.t.} \quad \Delta\theta_1 + \Delta\theta_2 = 0.$$

The KKT conditions (with multiplier μ) are

$$H\Delta\theta + b + \mu(1, 1)^\top = 0, \qquad \Delta\theta_1 + \Delta\theta_2 = 0.$$

Expanding coordinates yields

$$\frac{4}{3}\Delta\theta_1 - \frac{1}{3} + \mu = 0, \qquad \frac{2}{3}\Delta\theta_2 + \mu = 0, \qquad \Delta\theta_1 + \Delta\theta_2 = 0.$$

From the second and third equations $\mu = -\frac{2}{3}\Delta\theta_2$ and $\Delta\theta_2 = -\Delta\theta_1$. Substituting into the first gives $\frac{4}{3}\Delta\theta_1 - \frac{1}{3} - \frac{2}{3}\Delta\theta_2 = 0 \Rightarrow 6\Delta\theta_1 = 1$, hence

$$\Delta\theta_{\mathrm{DIF}} = (\tfrac{1}{6},\, -\tfrac{1}{6}) \tag{117}$$

which lies on the tangent subspace and thus closely preserves feasibility to first order.

A.5 Details in Constrained Linear Regression

We consider per-sample

$$\ell_i(\theta) = \tfrac{1}{2}\, (x_i^\top \theta - y_i)^2,$$

The constrained optimum $\bar{\theta}$ solves

$$\begin{aligned} \min_{\theta} \quad & \tfrac{1}{n} \sum_{i=1}^{n} \ell_i(\theta) \\ \text{s.t.} \quad & A_{\mathrm{eq}}\theta = b_{\mathrm{eq}}, \quad A_{\mathrm{ineq}}\theta \le b_{\mathrm{ineq}}. \end{aligned} \tag{118}$$

Assume we remove the data sample (x_i, y_i). We use

$$H = \nabla^2 L(\bar{\theta}) = \tfrac{1}{n} X^\top X, \qquad g_i = \nabla_\theta \ell_i(\bar{\theta}) = (x_i^\top \bar{\theta} - y_i)\, x_i.$$

Classical Influence Function (IF). Ignoring feasibility constraints, the first-order effect of removing sample (x_i, y_i) is

$$\Delta\theta_{\mathrm{IF}} \ = \ -\frac{1}{n}\, H^{-1}\, g_i. \tag{119}$$

Penalty-based IF. We approximate feasibility via a penalized surrogate

$$\tilde{L}(\theta) = \tfrac{1}{2n}\|X\theta - y\|_2^2 + \tfrac{\rho}{2}\|A_{\text{eq}}\theta - b_{\text{eq}}\|_2^2 + \sum_{j=1}^{m} k\left((a_j^\top \theta - b_j)_+\right)^3, \tag{120}$$

where $(t)_+ = \max\{t, 0\}$, $\rho, k > 0$, and $a_j^\top$ is the j-th row of A_{ineq}. The Hessian at $\bar{\theta}$ is

$$H_{\text{pen}} = \tfrac{1}{n}X^\top X + \rho\, A_{\text{eq}}^\top A_{\text{eq}} + \sum_{j\in\mathcal{A}} 6k\, t_j\, a_j a_j^\top, \qquad t_j = a_j^\top \bar{\theta} - b_j, \quad \mathcal{A} = \{j : t_j > 0\}. \tag{121}$$

The corresponding update is

$$\Delta\theta_{\text{penIF}} \;=\; -\frac{1}{n}\, H_{\text{pen}}^{-1}\, g_i. \tag{122}$$

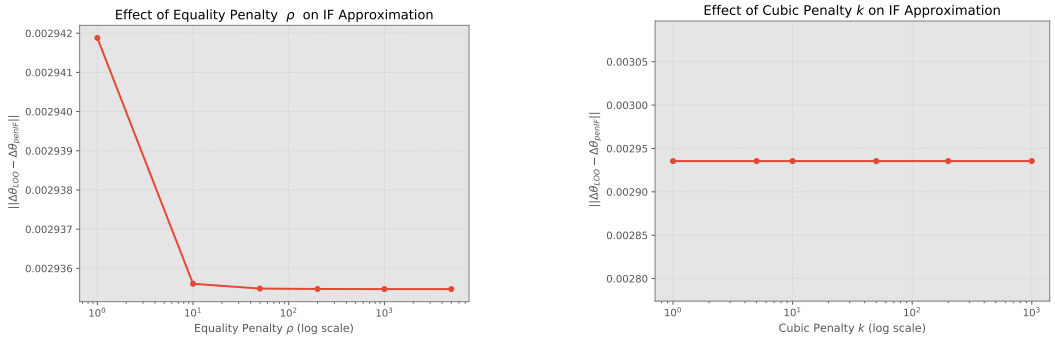
In the experiments shown in Figure 3, we set the penalty parameters to $\rho = 200$ and $k = 1000$. To justify this choice, we evaluated the discrepancy between Penalty-based IF estimation $\Delta\theta_{\text{penIF}}$ and leave-one-out retraining $\Delta\theta_{\text{LOO}}$:

$$\|\Delta\theta_{\text{penIF}} - \Delta\theta_{\text{LOO}}\|.$$

by varying both k and ρ.

Recall that $\bar{\theta}$ denotes the optimum of the constrained linear regression problem (118). At $\bar{\theta}$, all inequality constraints are strictly satisfied (i.e., $a_j^\top \bar{\theta} - b_j \le 0$ for every j), which implies that the active set is empty: $\mathcal{A} = \varnothing$. Consequently, the cubic penalty term contributes no curvature to the local Hessian H_{pen}, and therefore the value of k does not influence the penalty-based IF update $\Delta\theta_{\text{penIF}}$. This behavior is precisely reflected in Figure 6(b), where varying k produces essentially identical estimates.

In contrast, Figure 6(a) shows that increasing the equality–constraint penalty ρ steadily improves the approximation accuracy. The improvement becomes marginal once $\rho \ge 10$, after which the curve nearly plateaus. Based on this behavior, we adopt $\rho = 200$ in our experiments to ensure both stability and sufficient accuracy.



(a) (b)

Figure 6: Hyper-parameter sensitivity of the penalty-based influence function. (a) Effect of the equality penalty parameter ρ on the approximation error $\|\Delta\theta_{\text{penIF}} - \Delta\theta_{\text{LOO}}\|$. Increasing ρ improves the accuracy, but the gain becomes negligible once $\rho \ge 10$. (b) Effect of the cubic penalty parameter k on the cosine similarity between $\Delta\theta_{\text{penIF}}$ and $\Delta\theta_{\text{LOO}}$. Since the inequality constraint is non-active at the optimum $\bar{\theta}$, varying k results in almost no change.

Directional Influence Function (DIF). We enforce feasibility with a one-step constrained QP (linearized KKT):

$$\begin{aligned} \Delta\theta_{\text{DIF}} = \arg\min_{\Delta\theta\in\mathbb{R}^d} \quad & \tfrac{1}{2}\,\Delta\theta^\top H\,\Delta\theta \;-\; \tfrac{1}{n}\, g_i^\top \Delta\theta \\ \text{s.t.} \quad & A_{\text{eq}}\Delta\theta = 0, \\ & A_{\text{ineq}}\Delta\theta = 0 \;\; (\text{active constraints only}). \end{aligned} \tag{123}$$

A.6 Computational Complexity and Scalable Implementation of the QP Solver

A.6.1 Computational Complexity of the DIF QP

We recall that the quadratic program used to compute DIF takes the form

$$\min_{\omega} \quad b^\top \omega + \frac{1}{2}\,\omega^\top H_{\theta\theta}\omega \tag{124}$$

$$\text{s.t.} \quad g_j(\omega) := a_j^\top \omega + \eta_j \begin{cases} = 0, & j \in I_{\text{binding}}, \\ \le 0, & j \in I_{\text{non-binding}}, \end{cases} \tag{125}$$

where $H_{\theta\theta}$ is the Hessian of the Lagrangian with respect to θ at $(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})$,

$$b = \nabla_{\theta\varepsilon} L(\bar{\varepsilon}, \bar{\theta}, \bar{\lambda})\, \Delta\bar{\varepsilon},$$

and

$$a_j := \frac{1}{N_j} \sum_{i=1}^{N_j} \nabla_\theta \ell_j(z_i^{(j)}, \bar{\theta})$$

is the linearized constraint direction. Let $N_c = |I_{\text{binding}} \cup I_{\text{non-binding}}|$ denote the number of active or potentially active constraints. We collect the constraint directions into $A \in \mathbb{R}^{N_c \times d}$, whose j-th row is $a_j^\top$, and collect the constant terms into $\eta \in \mathbb{R}^{N_c}$.

Primal–dual interior-point formulation. For notational simplicity, we first describe the inequality part of the system; binding equality constraints are handled by the same KKT system without slack variables. Introducing slack variables $\zeta \in \mathbb{R}^{N_c}$ and dual variables $\xi \in \mathbb{R}^{N_c}$, the inequality constraints are written as

$$A\omega + \zeta + \eta = 0, \qquad \zeta > 0, \qquad \xi > 0.$$

With a logarithmic barrier parameter $\rho > 0$, the primal–dual residuals are

$$\begin{aligned} r_{\text{dual}} &= H_{\theta\theta}\omega + b + A^\top \xi, \\ r_{\text{pri}} &= A\omega + \zeta + \eta, \\ r_{\text{cent}} &= \Gamma\Lambda\mathbf{1} - \rho\mathbf{1}, \end{aligned}$$

where $\Gamma = \operatorname{diag}(\zeta)$ and $\Lambda = \operatorname{diag}(\xi)$.

Newton system. Each primal–dual iteration computes a Newton direction $(\Delta\omega, \Delta\xi, \Delta\zeta)$ by solving

$$\begin{bmatrix} H_{\theta\theta} & A^\top & 0 \\ A & 0 & I \\ 0 & \Gamma & \Lambda \end{bmatrix} \begin{bmatrix} \Delta\omega \\ \Delta\xi \\ \Delta\zeta \end{bmatrix} = - \begin{bmatrix} r_{\text{dual}} \\ r_{\text{pri}} \\ r_{\text{cent}} \end{bmatrix}. \tag{126}$$

Although the full KKT matrix has dimension $(d+2N_c)\times(d+2N_c)$, the slack block can be eliminated efficiently because both Γ and Λ are diagonal.

From the second and third block equations,

$$A\Delta\omega + \Delta\zeta = -r_{\text{pri}}, \qquad \Gamma\Delta\xi + \Lambda\Delta\zeta = -r_{\text{cent}},$$

we obtain

$$\Delta\zeta = -r_{\text{pri}} - A\Delta\omega,$$

and

$$\Delta\xi = \Gamma^{-1}\left(\Lambda A\Delta\omega - r_{\text{cent}} + \Lambda r_{\text{pri}}\right).$$

Substituting this expression into the first KKT block yields the Schur complement system

$$M_{\text{sc}}\Delta\omega = b_\omega, \tag{127}$$

where

$$M_{\text{sc}} = H_{\theta\theta} + A^\top\Gamma^{-1}\Lambda A,$$

and

$$b_\omega = -r_{\text{dual}} - A^\top\Gamma^{-1}\left(-r_{\text{cent}} + \Lambda r_{\text{pri}}\right).$$

After solving (127), the dual and slack directions are recovered by back-substitution:

$$\Delta\xi = \Gamma^{-1}\left(\Lambda A\Delta\omega - r_{\text{cent}} + \Lambda r_{\text{pri}}\right), \qquad \Delta\zeta = -r_{\text{pri}} - A\Delta\omega.$$

Matrix-free implementation. We do not explicitly form either the Hessian $H_{\theta\theta}$ or the Schur matrix M_{sc}. Instead, we solve (127) using a matrix-free Krylov method such as conjugate gradient. Each matrix–vector product with M_{sc} requires

$$M_{\mathrm{sc}} v = H_{\theta\theta} v + A^\top \Gamma^{-1} \Lambda A v.$$

The first term is computed by a Hessian–vector product (HVP), while the second term only requires products with A, $A^\top$, and diagonal matrices.

Computational complexity. Let C_{hvp} denote the cost of one HVP, which is comparable to a small number of backpropagation passes. Each Krylov iteration costs

$$\mathcal{O}\big(C_{\mathrm{hvp}} + N_c d\big),$$

where the $N_c d$ term comes from multiplying by A and $A^\top$. If K_{cg} Krylov iterations are used inside each Newton step and K_{pd} primal–dual iterations are performed, the total cost is

$$\mathcal{O}\left(K_{\mathrm{pd}} K_{\mathrm{cg}} \big(C_{\mathrm{hvp}} + N_c d\big)\right).$$

Since the number of active constraints N_c is typically much smaller than the parameter dimension d, the method scales linearly with the model size up to a small constraint-dependent factor. Importantly, the QP solver does not require forming the full Hessian or the full KKT matrix, making it suitable for large neural networks.

A.6.2 Scalable QP Solver with Structured Curvature Approximation

For large neural networks, solving the DIF QP with exact Hessian-vector products (HVP) can still be expensive. A scalable alternative is to replace $H_{\theta\theta}$ with a structured curvature approximation, such as K-FAC or EK-FAC, which enables efficient approximate IHVPs. We refer readers to [30, 15] for details on K-FAC/EK-FAC-based IHVP acceleration.

For large neural networks, solving the DIF QP with exact Hessian-vector products can still be expensive. To improve scalability, we replace the exact Hessian $H_{\theta\theta}$ in the QP solver with a structured curvature approximation. In particular, K-FAC and EK-FAC provide tractable layer-wise approximations to the Fisher/Gauss–Newton curvature and enable efficient iHVP. This yields a scalable approximate solver for the DIF QP.

Recall that, after eliminating the slack variables, the primal Newton direction is obtained from

$$\big(H_{\theta\theta} + A^\top D A\big)\, \Delta\omega = b_\omega, \tag{128}$$

where $D = \Gamma^{-1}\Lambda$, and $A \in \mathbb{R}^{N_c \times d}$ contains the linearized constraint directions. We approximate $H_{\theta\theta}$ by a structured curvature matrix G_{kfac}. Here, G_{kfac} denotes the K-FAC approximation to the curvature of the Lagrangian loss with respect to the model parameters. Specifically, K-FAC approximates each layer-wise curvature block by a Kronecker product of the input-activation covariance and the output-gradient covariance, enabling efficient inverse-vector products without forming the full Hessian. This gives, giving

$$\big(G_{\mathrm{kfac}} + A^\top D A\big)\, \Delta\omega = b_\omega. \tag{129}$$

Using the Woodbury identity, the solution can be written as

$$\Delta\omega = G_{\mathrm{kfac}}^{-1} b_\omega - G_{\mathrm{kfac}}^{-1} A^\top \left(D^{-1} + A G_{\mathrm{kfac}}^{-1} A^\top\right)^{-1} A G_{\mathrm{kfac}}^{-1} b_\omega. \tag{130}$$

This form avoids inverting a $d \times d$ matrix. Instead, it only requires applying G_{kfac}^{-1} to vectors and solving an $N_c \times N_c$ linear system. Since $N_c \ll d$ in our constrained learning settings, this substantially reduces the computational cost.

The dominant cost is forming

$$A G_{\mathrm{kfac}}^{-1} A^\top.$$

Because $A^\top$ has N_c columns, this requires applying the K-FAC inverse to each constraint direction. Thus, one Newton update costs approximately N_c applications of a K-FAC inverse-vector product, plus the cost of solving a small $N_c \times N_c$ system. If K_{pd} primal-dual iterations are used, the overall cost is

$$\mathcal{O}\big(K_{\mathrm{pd}} N_c\, C_{\mathrm{kfac}}^{-1}\big),$$

where C_{kfac}^{-1} denotes the cost of one K-FAC inverse-vector product. This provides a scalable alternative to exact HVP-based QP solving while preserving the constraint-adjusted structure of the DIF update.

Table 2: Performance comparison of different PINN models on the residual dataset after removal.

Model	MAE (Data Loss)	MAE (Physics Loss)	Relative L_2 Error (%)	Training Time (s)
Original Model	4.156	47.1×10^{-2}	21.65	327.06
Retrained Model	2.495	1.12×10^{-2}	14.24	331.50
Unlearned Model	2.832	0.37×10^{-2}	16.68	91.77

A.7 More experiments

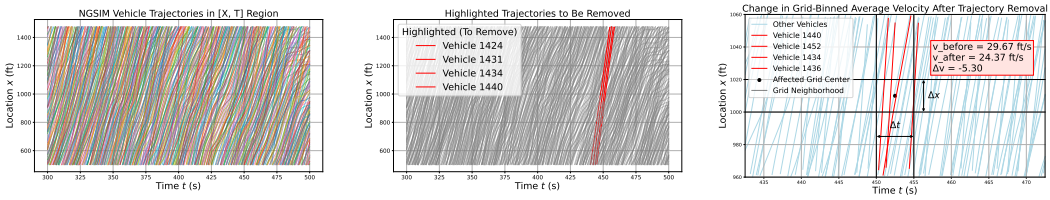
A.7.1 Machine unlearning of Physical informed Neural Network (PINN)

We further evaluate DIF on a machine unlearning task for a Physics-Informed Neural Network (PINN) model, following the formulation in [40]. The PINN is trained on the NGSIM vehicle-trajectory dataset [8] to estimate traffic velocity from observed trajectories, with a PDE constraint to enforce that the solution respects the traffic flow theory.

Let $u_\theta(x,t)$ be the neural-network approximation of the velocity field. The PINN training problem can be written as [4]

Figure 7 illustrates (i) the raw vehicle trajectories, (ii) the removal of a subset of the trajectory data to simulate unlearning, and (iii) the corresponding velocity field estimated by the PINN.

After removing 10% training data, we use DIF to estimate the change in the solution and obtain the updated model parameters. Table 2 presents the performance of three PINN models–Original, Retrained, and Unlearned–on the residual dataset after data removal. The Retrained Model achieves the best performance across all metrics, with the lowest data loss (MAE: 2.495), physics loss (MAE: 1.12×10^{-2}), and relative L_2 error (14.24%). The Unlearned Model shows slightly higher errors than the Retrained Model, but remains significantly better than the Original Model.



(a) (b) (c)

Figure 7: Illustration of trajectory removal and its impact on observed velocity. (a) All trajectories. (b) Subset of removed trajectories. (c) Change in average velocity in affected spatiotemporal bins due to trajectory removal.

A.7.2 Additional Large-Scale Validation on FairFace

We further evaluate DIF on a large-scale fairness-constrained learning problem. Specifically, we consider gender prediction on the FairFace dataset [22], which contains 108,501 face images from seven racial groups. We train a fairness-constrained CNN based on ResNet-50 and impose constraints to reduce risk disparity across racial groups.

[4]For simplicity of presentation in Appendix A.7.1, we reuse the notation (x,t,v) to denote the spatial coordinate, timestamp, and observed velocity in the NGSIM data. These symbols are independent of their usage in the main text.

$$\min_{\theta} \quad \frac{1}{N}\sum_{i=1}^{N}\big(u_\theta(x_i,t_i)-v_i\big)^2$$

subject to the PDE constraint (traffic-flow model)

$$\frac{\partial u_\theta}{\partial t}(x,t)+u_\theta(x,t)\frac{\partial u_\theta}{\partial x}(x,t)=0, \quad \forall (x,t)\in\Omega,$$

where (x_i,t_i,v_i) are the observed NGSIM trajectory data and Ω is the spatio-temporal domain.

Table 3: Loss change after removing the top 10% harmful training samples on the FairFace fairness-constrained ResNet-50 experiment. DIF identifies samples whose removal leads to the largest loss reduction on the target misclassified test example.

Method	DIF	TRAK	TracIn	Random
Loss change	**0.004427**	0.0037419	0.003235	0.0006043

We compare DIF with TRAK, TracIn, and random selection. Since TRAK and TracIn are originally designed for unconstrained learning, we adapt them to the constrained setting using the same counterfactual retraining protocol as in Section 5.2. For each method, we identify the top 10% most harmful training samples, remove them from the training set, and retrain the model. We then measure the loss change on a fixed misclassified test example after retraining. A larger loss reduction indicates that the selected training samples have stronger harmful influence on the target test example.

As shown in Table 3, DIF achieves the largest loss change after removing the selected harmful samples. This result suggests that DIF better captures the interaction between data influence and constraint satisfaction. In contrast, influence methods designed for unconstrained learning may overlook how a training sample affects the constrained solution through both the objective and the active fairness constraints, leading to less effective identification of harmful samples.

A.8 Details of Counterfactual Retraining Evaluation

We compare DIF with constrained TracIn and TRAK in the counterfactual retraining experiment. For each method, we compute an influence score for every training sample, remove the samples ranked as most harmful, and retrain the constrained CNN from the same training protocol. We evaluate the resulting loss change on the fixed misclassified test point used in the main experiment.

Constrained TracIn. Standard TracIn estimates the influence of a training point on a test point by accumulating gradient inner products along the training trajectory. In the constrained setting, we adapt TracIn by replacing the standard prediction loss with the Lagrangian loss used in the primal-dual training procedure. Specifically, during primal-dual optimization, we save checkpoints $\{\theta_{t_1}, \theta_{t_2}, \ldots, \theta_{t_k}\}$ and compute

$$\mathrm{TracInCP}(x_{\mathrm{train}}, x_{\mathrm{test}}) = \sum_{i=1}^{k} \delta_i \nabla_\theta \ell_{\mathrm{lag}}(\theta_{t_i}, x_{\mathrm{train}}) \cdot \nabla_\theta \ell_{\mathrm{lag}}(\theta_{t_i}, x_{\mathrm{test}}), \tag{131}$$

where ℓ_{lag} denotes the Lagrangian loss, θ_{t_i} is the model checkpoint at iteration t_i, and δ_i is the step size between checkpoints $i-1$ and i. This modification incorporates the constraint-adjusted training direction into the TracIn score.

TRAK. For TRAK, we apply the standard TRAK procedure to a penalized objective, where the fairness constraints are incorporated through a squared-hinge penalty. This gives TRAK access to constraint information through the surrogate loss, but it does not explicitly model the KKT-based sensitivity structure.

Removal protocol. For DIF, constrained TracIn, and TRAK, we rank training samples by their estimated harmfulness to the fixed misclassified test point. We then remove the top **2%** and **10%** most harmful samples and retrain the constrained CNN. Random removal is included as a reference baseline. The absolute loss reduction after retraining is reported in Figure 5.

Compute resources and societal impact. All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 GPU with 24GB memory. The constrained linear regression experiments were run on CPU, while the CNN, PINN, and FairFace experiments were run on GPU. DIF may improve interpretability, data debugging, and reliability analysis in constrained learning, especially when constraint violations are costly. DIF itself does not introduce additional high-risk capabilities beyond standard data-attribution methods. Its use should follow applicable privacy, data-governance requirements, and dataset terms of use.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:The abstract and introduction accurately summarize the main contributions, including the DIF formulation, the VI-based sensitivity analysis, and the empirical validation.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The final section discusses DIF's limitations, including its local nature and the Hessian-related computational costs inherited from classical influence functions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The main text explicitly states the regularity assumptions, including LICQ and SOSC, and the appendix provides complete proofs of all theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The appendix reports the hyperparameter choices, implementation details, and additional experimental settings needed to reproduce our results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper uses only publicly available datasets. The code is not publicly released at the time of submission, but we plan to release the implementation of DIF upon camera-ready acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The main text describes the experimental setup, and the appendix provides detailed DIF implementation settings, baselines' hyperparameter choices, and additional experimental details about influence functions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Our experimental results report variability across repeated trials, including mean and standard deviation for counterfactual retraining experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments were run on machines equipped with an NVIDIA RTX 4090 GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. It uses publicly available datasets and focuses on constrained data attribution without involving human-subject experiments, private data, or harmful deployment.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The paper discusses positive impacts, such as improving interpretability and reliability in constrained learning, as well as potential negative impacts, such as misuse for identifying sensitive or strategically important training samples.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk datasets or models, such as scraped datasets, pre-trained language models, or image generators.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The original sources of all datasets, models, and code assets used in the paper are properly credited, and their licenses and terms of use are respected.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new assets such as datasets, model checkpoints, benchmarks, or software packages.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The paper does not involve study participants or human-subject research, so participant risks, disclosure, and IRB approval are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as an important, original, or non-standard component of the core methodology, experiments, or analysis. Any use of LLMs was limited to writing, editing, or formatting assistance and did not affect the scientific content.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.