Generalized Fast Exact Conformalization

Diyang Li
Cornell University
diyang01@cs.cornell.edu

Abstract

Conformal prediction converts nearly any point estimator into a prediction interval under standard assumptions while ensuring valid coverage. However, the extensive computational demands of full conformal prediction are daunting in practice, as it necessitates a comprehensive number of trainings across the entire latent label space. Unfortunately, existing efforts to expedite conformalization often carry strong assumptions and are developed specifically for certain models, or they only offer approximate solution sets. To address this gap, we develop a method for fast exact conformalization of generalized statistical estimation. Our analysis reveals that the structure of the solution path is inherently piecewise smooth, and indicates that utilizing second-order information of difference equations suffices to approximate the entire solution spectrum arbitrarily. We provide a unified view that not only encompasses existing work but also attempts to offer geometric insights. Practically, our framework integrates seamlessly with well-studied numerical solvers. The significant speedups of our algorithm as compared to the existing standard methods are demonstrated across numerous benchmarks.

1 Introduction

In modern algorithmic practice, quantifying uncertainty is crucial for accurate and reliable model predictions. Conformal prediction [1] serves as a powerful statistical tool that leverages the observed data to construct prediction intervals containing the outcome with a predefined probability level. It enjoys model-free coverage guarantee regardless of the underlying distribution of the data. In recent years, conformal prediction has gained increasing attention from the community of machine learning [2–5], data mining [6–8] and computer vision [9, 10]. This growing interest is attributed to the attractive properties that it operates under the assumption of exchangeability, which is a weaker condition than independence and identical distribution, allowing for a wider range of applications in real-world scenarios where data may not meet strict statistical assumptions. Meanwhile, conformal prediction can be combined with almost any existing point estimators, even when the model is potentially misspecified [5].

While exhibits appealing properties, the application of conformal prediction often comes at a high computational cost [11, 12]. Kindly note that in this paper we refer to the full conformal prediction that does *not* discard training points as opposed to the split conformal prediction, as the latter involves only one single fitting. From a numerical perspective, when constructing a conformal prediction set, one needs to exhaustively search all points (potential candidates) in the label space, where for each point the learning model needs to be refitted and the conformity score needs to be re-calculated. In many scenarios like regression, the number of possible candidates is infinite as the latent label can take an uncountable number of possible values. Conventional conformal prediction works by a grid-search type method to loop over the label space [13, 14], which discretize the interval of interest and subsequently solve a sequence of individual optimization subproblems. To improve such brute-force approach, there have been many efforts in community that devoted to develop better algorithms for computing the prediction set. A natural idea is to generate the set of all solutions

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Representative related work, which are instances of *generalized parametric estimations*.

Model	Reference	Exact	Loss	Regularizer	Constrained	Path Structure
Least Squares	[15]	✓	Quadratic	$\setminus$	✗	Piecewise Linear
Ridge Regression	[16]	✓	Quadratic	$\|\mathbf{w}\|^2$	✗	Piecewise Linear
Empirical Risk Minimization	[17]	✗	Convex	Convex	✗	Piecewise Smooth
Elastic Net	[12]	✓	Quadratic	$\|\mathbf{w}\|_1$	✗	Piecewise Linear
Generalized Lasso Regression	[18]	✗	Convex	$\|\mathbf{w}\|_1$	✗	Piecewise Linear
General Formulation (**ours**)	Section 4	✓	PC^r	PC^r	✓	Piecewise Smooth

indexed by the latent label candidate using numerical continuation (*a.k.a.* homotopy) method, and we name this set of optimal solutions as (exact) *solution path*, as illustrated in Figure 1.

Despite extensive theoretical and empirical efforts, the understanding of conformalization path remains rather deficient. For some simple cases, closed-form characterizations of conformal prediction sets are available, such as k-nearest neighbors, least squares regression [15], and ridge regression [16] with quadratic loss. The study by [12] presents an exact solution path for conformalized Lasso and elastic net using their statistical property and ℓ_1-sparsity analysis. Investigations into more general objectives as discussed in [17, 18] incorporate linear interpolation to approximate the intrinsic piecewise smooth structure, yielding prediction sets lacking of finite-sample calibration. In other terms, [17, 18] offer only an upper bound for the approximation error and fail to control the degree of approximation relative to the optimal solutions. We summarize these relevant prior studies in Table 1. Given that existing exact algorithms are either tailored to a select few models, or turn out to be grid-search type approaches that take a very black-box approach, it prompts the following question:

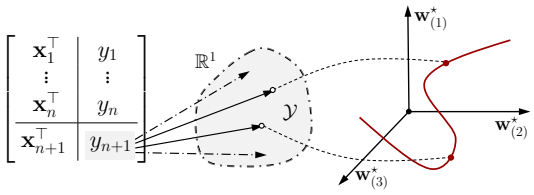

Figure 1: Diagram of setting on fast exact conformalization, where label candidate y_{n+1} should loop over the whole latent label space $\mathcal{Y}$ and each possible $y_{n+1} \in \mathbb{R}^1$ corresponds to one point on the (red) solution path. Conventional practice is to refit the model on each new y_{n+1} while we employ the path-following algorithm to obtain the whole solution spectrum within 1 execution.

Can we "open the black-box" by developing a methodology that better exploits the structure of the path?

We answer this question positively by introducing a differential equation perspective to analyze the ground truth solution path, which enables us to better reveal and exploit the fundamental path structure. This more profound understanding enables us to build more generalized conformalization algorithm and present improved computational guarantees.

1.1 Our contributions

The main contributions brought by this paper are summarized as follows.

Generalizable framework This study aims to extend the application of fast exact conformalization into generalized statistical estimation. We relax the assumptions by considering a cost function that is no longer globally differentiable, but rather *piecewise differentiable*, while introducing constraints on the weight vector to fit more statistical models. In our analysis, the Clarke subdifferential of the objective is derived, and the set of nonsmooth points of path can essentially be described as a level set of certain smooth functions. By adopting a reparameterization regime, our framework effectively broadens the scope of several existing algorithms in Table 1 and provides a unified view of them.

Theoretical insights We analyze the underlying structure of path through the first-order optimality conditions of the regularized problem, identifying sufficient conditions for a local path to be smooth around a given point and reveal that the structure of the path is inherently piecewise smooth. More precisely, if conditions are met, then the path is locally the projection of a higher-dimensional smooth manifold onto the optimization space, thus offering preliminary geometric intuitions. Our

investigation further suggests that leveraging the second-order information of difference equations can approximate the solution spectrum arbitrarily.

Practical efficiency Practically, our framework is not only straightforward to implement but also computationally efficient. With theoretical analysis, we present explicit expressions for the gradient flows of objective, which is homogenized and well aligned with standard forms used by mainstream numerical libraries for ordinary differential equations (ODE), thereby easing the programming efforts. When crossing potential kinks (or nonsmooth points), the computations are facilitated by boundary conditions pre-set in the numerical solver. Notably, our algorithm eliminates the need for extensive iterations for computing the entire solution spectrum, contrasting sharply with conventional baselines. In experiments, we demonstrate the significant computational speed advantages of our algorithm over existing baselines, without compromising on accuracy.

2 Background

Notation For a set $\mathcal{J} \subseteq \mathbb{R}^p$, we denote by $\mathrm{cl}(\mathcal{J})$ the closure and by $\mathrm{int}(\mathcal{J})$ the interior of $\mathcal{J}$ w.r.t. the natural topology on $\mathbb{R}^p$. Define $\mathcal{J}(i)$ the i-th element of $\mathcal{J}$ and $\mathrm{conv}(\mathcal{J})$ be the convex hull of $\mathcal{J}$, or $\mathrm{conv}(\mathcal{J}) = \{v : v = \sum_{i=1}^k \hat{\theta}_i u_i, u_i \in \mathcal{J}, \hat{\theta}_i \in \mathbb{R}^{>0}, \sum_{i=1}^k \hat{\theta}_i = 1\}$. The $\mathbf{w}_{(i)}$ is the i-th element of vector $\mathbf{w}$ and $\mathbb{I}(\cdot)$ is an indicator function. Let $\dot{z}(t)$ be the derivative $\frac{dz(t)}{dt}$ of function $z(t)$. Let $\mathbf{O}$ be the zero matrix and $\mathbb{P}\{\cdot\}$ be the event probability. The $\mathrm{sgn}(\cdot)$ is the sign function $\mathrm{sgn}(x) = \frac{x}{|x|}(x \neq 0)$ or 0 $(x = 0)$ that applied entrywise.

2.1 Model and assumptions

Given the dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^p$ is the sample (covariate) and y_i is the i-th label (response) live in label space $\mathcal{Y} \subseteq \mathbb{R}^1$. We consider the generalized parametric estimation problem

$$\begin{aligned} \mathbf{w}^\star \in \arg\min_{\mathbf{w}} \ & \sum_{i=1}^n L_i\left(y_i, \eta_{\mathbf{w}}(\mathbf{x}_i)\right) + \sum_{j=1}^m \lambda_j \Omega_j(\mathbf{w}) \\ s.t. \quad & g_i(\mathbf{w}) = 0, \quad 1 \leq i \leq r, \\ & h_j(\mathbf{w}) \leq 0, \quad 1 \leq j \leq s, \end{aligned} \tag{1}$$

where $\eta_{\mathbf{w}}$ is the model prediction function, L_i is the loss and Ω_j is the regularizer. The g_i, h_j are constraints on $\mathbf{w}$, and parameter $\lambda_j \in \mathbb{R}^{>0}$ controls the degree of regularity. In the following, we will define the piecewise differentiability and state main assumptions that used throughout this work.

Definition 1 (PC^r Function)**.** *Let $f : U \to \mathbb{R}$ be continuous on the open set $U \subseteq \mathbb{R}^p$ and $f_i : U \to \mathbb{R}$, $i \in \{1, \ldots, k\}$ be a set of r-times continuously differentiable (or C^r) functions for $r \in \mathbb{N} \cup \{\infty\}$. If $f(x) \in \{f_i(x)\}_{i \in \{1,\ldots,k\}}$ holds for all $x \in U$, then f is an r-times piecewise continuously differentiable (or PC^r) function. The $\{f_1, \ldots, f_k\}$ is a set of selection functions of f.*

When working with PC^r-functions in a local sense, it is useful to only consider the selection functions that have an impact on the local behavior around a given point.

Definition 2 (Essentially Active Set)**.** *Let $f : U \to \mathbb{R}$ be a PC^r function on the open set $U \subseteq \mathbb{R}^p$ with a set of selection functions $\{f_1, \ldots, f_k\}$. Denote $I_K = \{1, \ldots, k\}$. Then given a $x_1 \in U$, $I_f^a(x_1) \triangleq \{i \in I_K : f(x_1) = f_i(x_1)\}$ is called the active set at x_1, and the $I_f^e(x_1) \triangleq \{i \in I_K : x_1 \in \mathrm{cl}(\mathrm{int}(\{x_2 \in U : f(x_2) = f_i(x_2)\}))\}$ is the essentially active set at x_1.*

Assumption 1. *We assume that L_i and Ω_j in (1) are PC^r functions each with a set of selection functions $\bigcup_{k \in I_{L_i}^e} \{D_{L_i}^k\}$ and $\bigcup_{k \in I_{\Omega_j}^e} \{D_{\Omega_j}^k\}$, respectively.*

Assumption 2. *We assume that Ω_j and L_i are non-differentiable at $\mathbf{w}$ with multiple active selection functions, where $j \in \{1, \ldots, m\}$, $i \in \{1, \ldots, n+1\}$. We further assume that $I_{\Omega_j}^a(\mathbf{w}) \equiv I_{\Omega_j}^e(\mathbf{w}), I_{L_i}^a(\mathbf{w}) \equiv I_{L_i}^e(\mathbf{w})$ holds for all $\mathbf{w}$ considered and all Ω_j, L_i in the following.*

Assumption 2 ensures that all selection functions we consider are actually relevant for the representation of Ω_j, L_i. *i.e.*, it does not matter if we consider the active or the essentially active set in the underlying optimizing space, which allows for an easier representation of $D_{\Omega_j}^k$ and $D_{L_i}^k$.

2.2 Conformal prediction

Definition 3 (Symmetrical Algorithm)**.** *A deterministic algorithm* $A : (x_1, \ldots, x_n) \to A^\star$ *is symmetric if for any permutation* τ *of* $\{1, \ldots, n\}$, $A(x_1, \ldots, x_n) \overset{a.s.}{=} A(x_{\tau(1)}, \ldots, x_{\tau(n)})$.

Definition 4 (Conformity Score)**.** *The conformity score function* $\mathcal{A}$*, symmetric in its first* n *inputs, is defined as* $\mathcal{A}(\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_n; \vec{\mathbf{x}}_{n+1}) \colon \mathbb{R}^{(p+1)\times(n+1)} \to \mathbb{R}$*, where* $\vec{\mathbf{x}}_i \triangleq (\mathbf{x}_i, y_i)$.

Conformal prediction starts from a conventional model fitting stage, followed by the evaluation of conformity score and the construction of prediction set. The score function $\mathcal{A}$ serves as a measure of deviation or conformity, assessing the extent to which the new input $\mathbf{x}_{n+1}$ aligns with the previously fitted model. A higher conformity score indicates a better match between $\mathbf{x}_{n+1}$ and the model [1, 14]. For a new instance $\mathbf{x}_{n+1}$ where the prediction region is desired, the conformalization method operates by assigning a p-value to each latent $y_{n+1} \in \mathcal{Y}$, formalized as

$$\hat{p}_{y_{n+1}} = 1 - \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbb{I}\left(\mathcal{A}_i \geq \mathcal{A}_{n+1}\right), \tag{2}$$

where $\mathcal{A}_i \triangleq \mathcal{A}(\vec{\mathbf{x}}_1, .., \vec{\mathbf{x}}_{i-1}, \vec{\mathbf{x}}_{i+1}, .., \vec{\mathbf{x}}_{n+1}; \vec{\mathbf{x}}_i)$. Specifically, in density estimation, $\mathcal{A}$ is defined as $\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})$, where $\eta_{\mathbf{w}^\star}$ is the density function estimated from the augmented dataset. For regression tasks, $\mathcal{A}$ might be set as $-|y_{n+1} - \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})|$, with $\eta_{\mathbf{w}^\star}$ being the regression function trained by the dataset including $n+1$ samples. Under the assumption of exchangeability among the pairs $\{\mathbf{x}_i, y_i\}_{i=1}^{n+1}$, the $\hat{p}_{y_{n+1}}$ returned by (2) has been demonstrated to be statistically valid [13, 14]. To generate the prediction set $\Gamma(\cdot)$, one thresholds these p-values at a prescribed error level $\alpha \in (0, 1)$, resulting in

$$\Gamma(\mathbf{x}_{n+1}) = \left\{y_{n+1} : \hat{p}_{y_{n+1}} \geq \alpha\right\}. \tag{3}$$

Theorem 1. *[19] Suppose that* $\{\mathbf{x}_i, y_i\}_{i=1}^{n+1}$ *are exchangeable and the fitting algorithm* A *is symmetric. Conformal prediction applied on* $\{\mathbf{x}_i, y_i\}_{i=1}^{n} \cup \{\mathbf{x}_{n+1}\}$ *outputs a set* $\Gamma(\cdot)$ *such that*

$$\mathbb{P}\left\{y_{n+1}^\star \in \Gamma(\mathbf{x}_{n+1})\right\} \geq 1 - \alpha, \tag{4}$$

where $y_{n+1}^\star$ *is the ground truth* $(n+1)$*-th label.*

Theorem 1 (*a.k.a.* coverage guarantee) requires only exchangeability of input data and symmetry of the conformity score function, which are met by nearly all prevalent model fitting algorithms. In alignment with the analysis taken in prior research, we treat $y_{n+1} := y_{n+1}(z)$ as a function of scalar variable z. We utilize it to facilitate the traversal of y_{n+1} across the entire label space $\mathcal{Y}$, and compute the homotopy solution path $\{\mathbf{w}^\star(z) : z_{\min} \leq z \leq z_{\max}\}$ (also shown in Figure 1). We assume that $y_{n+1}(\cdot) \colon [z_{\min}, z_{\max}] \to \mathcal{Y}$ is continuously differentiable in terms of z for simplicity.

3 Main results

We present our main results for fast exact conformalization. The discussions here focus on positive z,[1] but the derivation extends easily to include negative values as well, which will be discussed later.

3.1 Surrogate function

Lemma 1. *Let* $f \colon U \to \mathbb{R}$ *be a* PC^r*-function on the open set* $U \subseteq \mathbb{R}^p$ *and let* C^r*-functions* $\{f_1, \ldots, f_k\}$ *be a set of selection functions of* f*. Then for any* $x \in U$*, there exists an open neighborhood* $U' \subseteq U$ *of* x *on which* f *is also a continuous selection of* $\{f_i \colon i \in I_f^e(x)\}$.

Drawing insights from Lemma 1, we minimize the surrogate function of (1) as

$$\min_{\mathbf{w}\in\mathbb{R}^p} \mathcal{E}_z(\mathbf{w}) := \sum_{i=1}^{n+1} L_i\left(y_i, \eta_{\mathbf{w}}(\mathbf{x}_i)\right) + \sum_{j=1}^{m} \lambda_j \Omega_j(\mathbf{w}) + \rho \sum_{i=1}^{r} |g_i(\mathbf{w})| + \rho \sum_{j=1}^{s} \max\left\{0, h_j(\mathbf{w})\right\}, \tag{5}$$

[1] For technicality reasons, we enforce $z > 0$ even the limit $\mathbf{w}^\star(0^+) \triangleq \lim_{z\to 0^+} \mathbf{w}^\star(z)$ might be well-defined.

where $\rho \in \mathbb{R}^{>0}$ and $y_{n+1} = y_{n+1}(z)$. This definition of $\mathcal{E}_z(\mathbf{w})$ is meaningful regardless of whether the contributing functions are convex. Denote model estimation terms $\sum_i L_i + \sum_j \lambda_j \Omega_j$ as $\mathbb{L}_M(\mathbf{w}|z)$, and the minimizer of (5) as $\mathbf{w}^\star(z)$. It is interesting to compare $\mathcal{E}_z(\mathbf{w})$ to the Lagrangian function

$$\mathcal{L}_z(\mathbf{w}) := \mathbb{L}_M(\mathbf{w}|z) + \sum_{i=1}^{r} \tilde{\lambda}_i g_i(\mathbf{w}) + \sum_{j=1}^{s} \tilde{\mu}_j h_j(\mathbf{w}), \tag{6}$$

which captures the behavior of $\mathbb{L}_M(\cdot)$ near the optimum. At a constrained minimum $\mathbf{w}^\star$, the Lagrangian satisfies the stationarity condition $\nabla\mathcal{L}(\mathbf{w}^\star) = \mathbf{0}$; its inequality multipliers $\tilde{\mu}_j$ are nonnegative and satisfy the complementary slackness $\tilde{\mu}_j h_j(\mathbf{w}^\star) = 0$. In the penalized (5), one usually takes

$$\rho > \max\{|\tilde{\lambda}_1|, \ldots, |\tilde{\lambda}_r|, \tilde{\mu}_1, \ldots, \tilde{\mu}_s\}, \tag{7}$$

which creates the favorable circumstances: **(i)** $\mathcal{L}_z(\mathbf{w}) \leq \mathcal{E}_z(\mathbf{w})$ for all $\mathbf{w}$, **(ii)** $\mathcal{L}_z(\mathbf{w}) \leq \mathbb{L}_M(\mathbf{w}|z) = \mathcal{E}_z(\mathbf{w})$ for all feasible $\mathbf{w}$, **(iii)** $\mathcal{L}_z(\mathbf{w}^\star) = \mathbb{L}_M(\mathbf{w}^\star|z) = \mathcal{E}_z(\mathbf{w}^\star)$ with profound consequences.

Theorem 2. *Our surrogate function $\mathcal{E}_z(\mathbf{w})$ is increasing in ρ. Furthermore, $\mathcal{E}_z(\mathbf{w})$ is strictly convex for one $\rho > 0$ if and only if it is strictly convex for all $\rho > 0$. Likewise, it is coercive for one $\rho > 0$ if and only if it is coercive for all $\rho > 0$. Finally, if $\mathbb{L}_M(\cdot)$ is strictly convex (or coercive), then $\mathcal{E}_z(\mathbf{w})$ for all ρ are strictly convex (or coercive).*

Given Theorem 2, several classical results [20] state that minimizing $\mathcal{E}_z(\mathbf{w})$ is effective in minimizing $\mathbb{L}_M(\cdot)$ subject to the constraints if we choose ρ by (7).

3.2 Conformal path characterization

Definition 5 (Clarke Subdifferential)**.** *For continuous function $f : U \to \mathbb{R}$ on the open set $U \subseteq \mathbb{R}^p$, let $\Theta \subseteq U$ be the set of points in which f is not differentiable. The Clarke subdifferential of f at $x \in U$ is defined as*

$$\partial f(x) \triangleq \mathrm{conv}(\{\phi \in \mathbb{R}^p : \exists\{x_i\}_{i=1}^{\infty} \in \mathbb{R}^p \backslash \Theta \text{ with } \lim_{i\to\infty} x_i = x, \lim_{i\to\infty} \nabla f(x_i) = \phi\}).$$

Specifically, if f is continuously differentiable in x, $\partial f(x) = \{\nabla f(x)\}$. While PC^r-functions are generally non-smooth, we can use the Clarke subdifferential to obtain first-order optimality of (5),

$$\sum_{i=1}^{n} \sum_{k\in I^a_{L_i}(\mathbf{w}^\star)} \hat{\theta}^k_{L_i}(\mathbf{w}^\star) \nabla D^k_{L_i}(y_i, \eta_{\mathbf{w}^\star}(\mathbf{x}_i)) + \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star) \nabla D^k_{L_{n+1}}(y_{n+1}(z), \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})) \triangleq \mathbf{D}'(\mathbf{w}^\star),$$

$$\rho \sum_{i=1}^{r} \hat{\theta}_{g_i} \nabla g_i(\mathbf{w}^\star) + \rho \sum_{j=1}^{s} \hat{\theta}_{h_j} \nabla h_j(\mathbf{w}^\star) + \sum_{j=1}^{m} \sum_{k\in I^a_{\Omega_j}(\mathbf{w}^\star)} \lambda_j \hat{\theta}^k_{\Omega_j}(\mathbf{w}^\star) \nabla D^k_{\Omega_j}(\mathbf{w}^\star) + \mathbf{D}'(\mathbf{w}^\star) = \mathbf{0}, \tag{8}$$

where $\hat{\theta}^k_{\Omega_j}, \hat{\theta}^k_{L_i}$ is the k-th auxiliary parameter for convex hull of each Ω_j, L_i. The (8) is accompanied by the active sets conditions and the subdifferentials conditions, rewritten in detail as

$$\begin{aligned}
&D^k_{L_i}(\mathbf{w}^\star) - D^{l_i}_{L_i}(\mathbf{w}^\star) = 0,\ \forall k \in I^a_{L_i}(\mathbf{w}^\star)\backslash\{l_i\},\ \forall i \in \bar{I}^e_{L_i} \\
&D^k_{\Omega_j}(\mathbf{w}^\star) - D^{r_j}_{\Omega_j}(\mathbf{w}^\star) = 0,\ \forall k \in I^a_{\Omega_j}(\mathbf{w}^\star)\backslash\{r_j\},\ \forall j \in \bar{I}^e_{\Omega_j} \\
&\sum_{k\in I^a_{L_i}(\mathbf{w}^\star)} \hat{\theta}^k_{L_i}(\mathbf{w}^\star) - 1 = 0,\ \hat{\theta}^k_{L_i}(\mathbf{w}^\star) \geq 0, 1 \leq i \leq n+1 \\
&\sum_{k\in I^a_{\Omega_j}(\mathbf{w}^\star)} \hat{\theta}^k_{\Omega_j}(\mathbf{w}^\star) - 1 = 0,\ \hat{\theta}^k_{\Omega_j}(\mathbf{w}^\star) \geq 0,\ 1 \leq j \leq m
\end{aligned} \tag{9}$$

where r_j, l_i is randomly selected from $I^a_{\Omega_j}, I^a_{L_i}$ and being fixed, with coefficients satisfying $\hat{\theta}_{g_i} \in \begin{cases} \{-1\} & g_i(\mathbf{w}) < 0 \\ [-1,1] & g_i(\mathbf{w}) = 0 \\ \{1\} & g_i(\mathbf{w}) > 0 \end{cases}, \hat{\theta}_{h_j} \in \begin{cases} \{0\} & h_j(\mathbf{w}) < 0 \\ [0,1] & h_j(\mathbf{w}) = 0 \\ \{1\} & h_j(\mathbf{w}) > 0 \end{cases}$. In this work, we specialize to the case where the

constraint functions g_i $(1 \le i \le r)$ and h_j $(1 \le j \le s)$ are *affine*, *i.e.*, the gradients $\nabla g_i(\mathbf{w}), \nabla h_j(\mathbf{w})$ are constant.[2] We define g_i and h_j as constraint residuals $g_i(\mathbf{w}) := \mathbf{v}_i^\top \mathbf{w} - d_i$, $h_j(\mathbf{w}) := \boldsymbol{\omega}_j^\top \mathbf{w} - e_j$.

We keep track of the following index sets determined by signs of constraint residuals:

$$\begin{aligned}
&\mathcal{N}_{\mathrm{E}} = \{i : g_i(\mathbf{w}) = \mathbf{v}_i^\top \mathbf{w} - d_i < 0\}, \mathcal{Z}_{\mathrm{E}} = \{i : g_i(\mathbf{w}) = \mathbf{v}_i^\top \mathbf{w} - d_i = 0\}, \mathcal{P}_{\mathrm{E}} = \{i : g_i(\mathbf{w}) = \mathbf{v}_i^\top \mathbf{w} - d_i > 0\}, \\
&\mathcal{N}_{\mathrm{I}} = \{j : h_j(\mathbf{w}) = \boldsymbol{\omega}_j^\top \mathbf{w} - e_j < 0\}, \mathcal{Z}_{\mathrm{I}} = \{j : h_j(\mathbf{w}) = \boldsymbol{\omega}_j^\top \mathbf{w} - e_j = 0\}, \mathcal{P}_{\mathrm{I}} = \{j : h_j(\mathbf{w}) = \boldsymbol{\omega}_j^\top \mathbf{w} - e_j > 0\}.
\end{aligned} \tag{10}$$

To characterize the path, we further introduce a *reparameterization* in terms of an auxiliary variable $t \ge 0$ (thought of as *time*), whereby for a given $\mathbb{T} > 0$ we introduce functions $z(\cdot) : [0, \mathbb{T}] \to [z_{\min}, z_{\max}]$ and $\xi(\cdot) : [z_{\min}, z_{\max}] \to \mathbb{R}$ such that: $\xi(\cdot)$ is Lipschitz, $z(\cdot)$ is differentiable on $(0, \mathbb{T})$, and we have $\dot{z} = \xi(z(t))$ for all $t \in (0, \mathbb{T})$. In a slight abuse of notation, we define the path w.r.t. t as $\left\{ \mathbf{w}^\star(t) \triangleq \mathbf{w}^\star(z(t)) : t \in [0, \mathbb{T}] \right\}$. To enhance structural clarity, we denote $n_{\mathcal{Z}} = |\mathcal{Z}_{\mathrm{E}} \cup \mathcal{Z}_{\mathrm{I}}|$ and represent the certain matrix inversion as

$$\begin{bmatrix} \tilde{\mathbf{H}}(\mathbf{w}^\star|z) & \mathbf{U}_{\mathcal{Z}}^\top \\ \mathbf{U}_{\mathcal{Z}} & \mathbf{O}_{n_{\mathcal{Z}} \times n_{\mathcal{Z}}} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{P}(\mathbf{w}^\star|z) & \mathbf{Q}(\mathbf{w}^\star|z) \\ \mathbf{Q}^\top(\mathbf{w}^\star|z) & \mathbf{R} \end{bmatrix}, \tag{11}$$

where the rows of matrix $\mathbf{U}_{\mathcal{Z}}$ are the constant differentials, $\mathbf{v}_i^\top$ for $i \in \mathcal{Z}_{\mathrm{E}}$ and $\boldsymbol{\omega}_j^\top$ for $j \in \mathcal{Z}_{\mathrm{I}}$, and

$$\begin{aligned}
\tilde{\mathbf{H}}(\mathbf{w}^\star|z) \triangleq & \sum_{i=1}^{n} \sum_{k \in I^a_{L_i}(\mathbf{w}^\star)} \hat{\theta}^k_{L_i}(\mathbf{w}^\star) \nabla^2 D^k_{L_i}(y_i, \eta_{\mathbf{w}^\star}(\mathbf{x}_i)) + \sum_{j=1}^{m} \sum_{k \in I^a_{\Omega_j}(\mathbf{w}^\star)} \lambda_j \hat{\theta}^k_{\Omega_j}(\mathbf{w}^\star) \nabla^2 D^k_{\Omega_j}(\mathbf{w}^\star) \\
& + \sum_{k \in I^a_{L_{n+1}}(\mathbf{w}^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star) \nabla^2 D^k_{L_{n+1}}(y_{n+1}(z), \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})).
\end{aligned} \tag{12}$$

Now we are fully equipped to unveil the structure of homotopy path, as shown in next subsection.

3.3 Piecewise smooth structure & Unified view

To describe the structure, we define the kink (non-smooth point) as the point that $\hat{\theta}^k_{\Omega_j}$, $\hat{\theta}^k_{L_i}$ hit the restriction bound in (9), or set in (10) is violated so the entire structure changes.

Theorem 3. *Given an optimum $(\mathbf{w}_0^\star, z_0)$ at t_0, and assume that t_0 is not a kink, then there exists an open neighborhood of t_0 such that $\mathbf{w}^\star(t)$ is a C^1 function of t and satisfies the following autonomous system*

$$\dot{\mathbf{w}}^\star(t) \triangleq \Upsilon(\mathbf{w}^\star, z) = - \sum_{k \in I^a_{L_{n+1}}(\mathbf{w}^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\, \xi(z) \cdot \mathbf{P}(\mathbf{w}^\star|z) \left[\partial_z \nabla D^k_{L_{n+1}}(y_{n+1}(z), \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})) \right], \tag{13}$$

where $\mathbf{P}(\mathbf{w}^\star|z)$ is obtained from the definition in (11), and the $\hat{\theta}_{L_i}$, $\hat{\theta}_{\Omega_j}$ can be solved from (8).

The (13) offers an explicit gradient flow of the solution path in underlying optimization space. Solving this ODE numerically would recover all the optimal parametric solutions, which provide the conformity scores and one can obtain the required p-values (2) that used for exact conformal prediction via (3).

Theorem 4. *The optimality solution $(\mathbf{w}^\star(t), z(t))$ has a unique trajectory for $t \in (0, \mathbb{T})$. The $\mathbf{w}^\star(t)$ is continuous if selections $D^k_{L_i}(\cdot)$, $D^k_{\Omega_j}(\cdot)$ are convex. Furthermore, if gradients of constraints $\{\nabla g_i(\mathbf{w}) : \nabla g_i(\mathbf{w}) = \mathbf{0}\} \cup \{\nabla h_j(\mathbf{w}) : \nabla h_j(\mathbf{w}) = \mathbf{0}\}$ are affinely independent at the solution $\mathbf{w}^\star(z)$ over an open neighborhood of z, then the coefficient paths $\hat{\theta}_{g_i}$, $\hat{\theta}_{h_j}$ are unique and continuous at z.*

Theorem 5. *On a optimality path with set configuration (10), the coefficients for constraints satisfies*

$$\boldsymbol{r}_{\mathcal{Z}} \triangleq \begin{bmatrix} \hat{\theta}_{\mathcal{Z}_E}(\mathbf{w}^\star) \\ \hat{\theta}_{\mathcal{Z}_I}(\mathbf{w}^\star) \end{bmatrix} = -\mathbf{Q}(\mathbf{w}^\star) \left[\frac{1}{\rho} \mathbf{D}'(\mathbf{w}^\star) + \mathbf{u}_{\bar{\mathcal{Z}}}^\top \right], \tag{14}$$

[2] In principle a similar algorithm can be developed for the general convex constraint where the h_j are relaxed to convex, but that is beyond the scope of current paper.

where $\mathbf{Q}(\mathbf{w}^\star)$ *is from (11), and*

$$\mathbf{u}_{\bar{\mathcal{Z}}}^\top := -\sum_{i\in\mathcal{N}_E}\mathbf{v}_i + \sum_{i\in\mathcal{P}_E}\mathbf{v}_i + \sum_{j\in\mathcal{P}_I}\boldsymbol{\omega}_j.$$

Although Theorem 3, 4, and 5 are highly technical and may difficult to grasp on first glance, they lay the groundwork for practical application, as illustrated later in Section 4. Specifically, Theorem 4 ensures continuity, which is fundamental for the numerical ODE solvers. Theorem 5, on the other hand, provides a rule for handling constraints.

Theorem 6. *Suppose that* $D_{L_i}^k(\cdot)$ *is* μ*-strongly convex for* $\mu \geq 0$*,* $D_{\Omega_j}^k(\cdot)$ *is* σ*-strongly convex for* $\sigma > 0$*, and* $\mathbf{P}(\mathbf{w}^\star|z)$*,* $\partial_z D_{L_{n+1}}^k(\cdot)$*,* $\partial_z \nabla D_{L_{n+1}}^k(\cdot)$ *are all locally* ℓ*-Lipschitz continuous. Suppose further that* $\xi(\cdot)$ *is Lipschitz continuous on* $[z_{\min}, z_{\max}]$ *and satisfies* $|\xi(z)| \leq \bar{C}$ *for all* $z \in [z_{\min}, z_{\max}]$*. Then, it holds that* $\Upsilon(\cdot,\cdot)$ *defined in (13) is uniformly* ℓ_Υ*-Lipschitz continuous with* $\ell_\Upsilon = \bar{C}\ell^2 + \frac{2\bar{C}\ell}{(n+1)\mu+\sum_{j=1}^m \lambda_j \sigma}$ *for any* $z \in [z_{\min}, z_{\max}]$ *when the active selections* I_L*,* I_Ω *are fixed.*

The essence of Theorem 6 lies in its suggestion that the dynamics of $\mathbf{w}^\star(t)$ is piecewise continuous, *i.e.*, $\mathbf{w}^\star(t)$ maintains smoothness between two adjacent kinks. By considering specific choices of $\xi(\cdot)$ and $y_{n+1}(z)$, our system (13) generalizes some previously studied methodologies in fast conformalization. First, consider the scenario with an equally spaced discretization of the interval $[0, \mathbb{T}]$, namely $t_k = k \cdot h'$ for some fixed step-size $h' > 0$. Thus, the sequence $z_k := z(t_k)$ is approximately given by $z_{k+1} \approx z_k + h' \cdot \xi(z_k)$. Intuitively, the choice of $\xi(\cdot)$ controls the dynamic of $z(\cdot)$ and generalizes some previously considered sequences $\{z_k\}$ for the problem (5). For example, letting $\xi(z) := 1$ we recover the arithmetic sequence in [12] and letting $\xi(z) := -z$ we recover the geometric sequence in [17].

4 Fast conformalization algorithm

The complete algorithm on the fast exact conformalization is outlined in the plate referred as Algorithm 1.

Algorithm 1 Fast Exact Conformalization Algorithm

Input: Training data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, new covariate $\mathbf{x}_{n+1}$, range $[z_{\min}, z_{\max}]$, initial solution $\mathbf{w}_0^\star$, regularization strength $\{\lambda_j\}_{j=1}^m$, miscoverage level $\alpha \in (0,1)$.

\\ `Full Path Generation`
$z \leftarrow 0$, set $\mathcal{N}_\mathrm{E}$, $\mathcal{P}_\mathrm{E}$, $\mathcal{P}_\mathrm{I}$ and all I_L, I_Ω by $\mathbf{w}_0^\star$.
while $0 \leq z \leq z_{\max}$ **do**
  while partitions $\mathcal{N}_\mathrm{E}$, $\mathcal{P}_\mathrm{E}$, $\mathcal{P}_\mathrm{I}$, I_L, I_Ω are met **do**
    Calculate $\tilde{\mathbf{H}}(\mathbf{w}^\star|z)$, $\boldsymbol{r}_{\bar{\mathcal{Z}}}$ as in (12), (14).
    Solve ODE system (13).
  end while
  Update $\mathcal{N}_\mathrm{E}$, $\mathcal{P}_\mathrm{E}$, $\mathcal{P}_\mathrm{I}$, I_L, I_Ω by index violator(s).
end while
$z \leftarrow 0$.
while $z_{\min} \leq z \leq 0$ **do**
  Repeat the above procedure analogously for negative values of z, obtaining $\{\mathbf{w}^\star(z) : z_{\min} \leq z \leq 0\}$.
end while
\\ `Conformal Set Generation`
for $i = 1$ to $n+1$ **do**
  Calculate conformity score path $\mathcal{A}_i$ for i-th sample.
end for
Calculate the path of $\hat{p}_{y_{n+1}}$ by (2).
$\Gamma(\mathbf{x}_{n+1}) \leftarrow \{y_{n+1} : \hat{p}_{y_{n+1}} \geq \alpha\}$.

Output: Conformal prediction set $\Gamma(\mathbf{x}_{n+1})$.

Table 2: Numerical results for average empirical coverage, average length of the conformal prediction set, and the *total* number of kinks observed. One standard error is given in the parenthesis following the average number.

Dataset	Model (Parameter)	Grid1/Grid2		SCP		Exact		# Kinks
		Coverage	Length	Coverage	Length	Coverage	Length	
fried	NLS(\)	0.81(0.003)	5.82(0.53)	0.81(0.007)	5.55(0.94)	0.80(0.009)	5.30(0.72)	6
cadata	NLS(\)	0.82(0.007)	5.54(0.12)	0.82(0.007)	5.63(0.94)	0.80(0.003)	5.35(0.02)	2
delta	NLS(\)	0.81(0.008)	5.63(0.16)	0.81(0.002)	5.79(0.53)	0.80(0.002)	5.26(0.27)	5
cadata	GFM($\lambda_1 = 0.03$)	0.82(0.001)	11.64(0.13)	0.81(0.004)	11.38(0.36)	0.80(0.003)	11.34(0.27)	7
cadata	GFM($\lambda_1 = 0.02$)	0.83(0.008)	10.72(0.91)	0.81(0.004)	10.39(0.97)	0.80(0.001)	10.13(0.21)	5
elevator	GFM($\lambda_1 = 0.03$)	0.83(0.003)	11.09(0.58)	0.81(0.007)	11.34(0.97)	0.80(0.002)	10.68(0.81)	9
fried	IGR($\lambda_1 = 0.1, \lambda_2 = 0.02$)	0.81(0.002)	19.14(0.78)	0.81(0.001)	19.69(0.96)	0.81(0.002)	19.05(0.19)	12
cadata	IGR($\lambda_1 = 0.1, \lambda_2 = 0.02$)	0.82(0.003)	19.99(0.33)	0.80(0.002)	19.16(0.29)	0.80(0.002)	19.14(0.82)	13
elevator	IGR($\lambda_1 = 0.2, \lambda_2 = 0.05$)	0.82(0.003)	19.99(0.85)	0.81(0.001)	19.53(0.53)	0.81(0.005)	19.48(0.75)	7

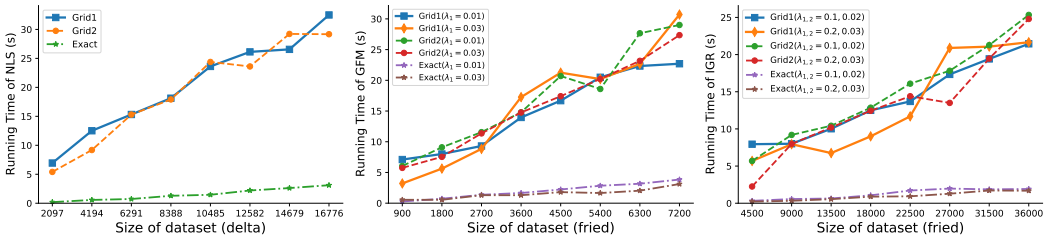

Figure 2: The running time under different sizes of datasets.

4.1 On solving (11)

To avoid the direct computation of matrix inversion (11), our practical algorithm involves sweep operator [22, 23]. Suppose $\mathbf{A}$ is an $p \times p$ symmetric matrix, sweeping on the k-th diagonal entry $A_{kk} \neq 0$ of $\mathbf{A}$ results in a matrix $\widehat{\mathbf{A}}$ with entries

$$\hat{A}_{kk} = -\frac{1}{A_{kk}}, \quad \hat{A}_{ik} = \frac{A_{ik}}{A_{kk}} \, (i \neq k), \quad \hat{A}_{kj} = \frac{A_{kj}}{A_{kk}} \, (j \neq k), \quad \hat{A}_{ij} = A_{ij} - \frac{A_{ik}A_{kj}}{A_{kk}} \, (i, j \neq k). \tag{15}$$

Since the sweeping (15) preserves symmetry, all operations can be performed solely on either the lower or upper-triangular part of $\mathbf{A}$ to ease the computational burden [24]. To begin with, we initiate with a sweeping tableau as $\left[\begin{array}{c|c} \tilde{\mathbf{H}}(\mathbf{w}^\star|z) & * \\ \hline \mathbf{U}_{\mathcal{Z}} & \mathbf{O} \end{array}\right]$, and further sweeping of diagonal entries of block $\tilde{\mathbf{H}}$ yields $\mathbf{M} \triangleq \left[\begin{array}{c|c} \mathbf{M}_{11} & * \\ \hline \mathbf{M}_{21} & \mathbf{M}_{22} \end{array}\right]$. Then we reinitialize our new tableau in the form of $\left[\begin{array}{c|c} -\mathbf{M}_{22} & \mathbf{M}_{21} \\ \hline * & -\mathbf{M}_{11} \end{array}\right]$, and further sweeping of diagonal entries of block $\mathbf{M}_{22}$ makes $\left[\begin{array}{c|c} \mathbf{R} & \mathbf{Q}^\top(\mathbf{w}^\star|z) \\ \hline * & \mathbf{P}(\mathbf{w}^\star|z) \end{array}\right]$. Compared to direct inversion, it also decreases a $\mathcal{O}(p^2 + n_{\mathcal{Z}}^2)$ storage space.

4.2 Efficiency

Algorithm 1 shows a very favourable behavior empirically, and converges remarkably faster than the standard grid-search type algorithm. We argue that this observation is actually quite natural. Indeed, our algorithm can follow the ground truth solution path, and the numerical integration process is fully deterministic, which avoids large fluctuations between the iteration steps like stochastic gradient descent. Therefore, the solving process of Algorithm 1 exhibits a more stable character being completely *deterministic* and has no extensive loops, which explains the much faster convergences observed in practice. In contrast, the standard grid-search type method would cost N-times the original batch iterative algorithms, where N is the number of grid points.

5 Numerical experiments

We provide experimental results on real-world benchmarks to validate our derived algorithm. All experiments presented in this study were conducted on a workstation running the Ubuntu 18.04 operating system, equipped with Intel Xeon `Gold 5218R` CPU$\times$64 and 60.9 GB of RAM. We integrate a system of ordinary differential equations using `lsoda` from the `FORTRAN` library, where an

interface for `SciPy` is available using the `odepack`. The concrete parameter settings of ODE solver are shown in the Table 3, wherein the numerical solver exploit the Runge-Kutta method of order 4 or 5. The parameterizers are set to $y_{n+1}(z) = 4z$, $\xi(z) = 1$, respectively.

Table 3: List of the key parameters in the numerical solver.

Parameters	Descriptions	Values
`rtol`	allowed relative error in the solution	`1e-6`
`atol`	allowed absolute error in the solution	`1.49e-8`
`tcrit`	vector of critical points	*set by known / explicit kinks*
`h0`	initial step size for the integration	`0.02*`$(z_{max} - z_{min})$
`hmax`	maximum absolute step size allowed	`0.1*`$(z_{max} - z_{min})$
`hmin`	minimum absolute step size allowed	`1e-7`
`mxstep`	maximum number of steps allowed for each point	`400`
`mxordn`	maximum order to be allowed for the non-stiff method	`12`
`mxords`	maximum order to be allowed for the stiff method.	`5`
`ixpr`	extra printing at method switches	`True`
`mxhnil`	maximum number of messages printed	`15`
`tfirst`	the required order of the first two arguments	`True`
`full_output`	return a dictionary of optional outputs	`True`

Model For evaluation, here we employ 3 specific forms of (1), *i.e.*, Nonnegative Least Squares (`NLS`) [25], Graph-guided Fused Model (`GFM`) [26], and Inverse Gaussian Regression (`IGR`) [27].

Conformalization A conformal prediction set with target coverage level 0.8 ($\alpha = 0.2$) is calculated for each sample in testing set using each of the 4 methods, *i.e.*, the standard grid-point evaluation method (`Grid1`) [1], grid-point method with warm-restart strategy (`Grid2`) [28], the split conformal prediction method (`SCP`) [4], and our exact conformalization method in Algorithm 1 (`Exact`). We use the conformity score function $\mathcal{A}_i = -|y_i - \eta_{\mathbf{w}^\star}(\mathbf{x}_i)|$. Conventionally, the interval $[y_{n+1}^{\min}, y_{n+1}^{\max}]$ (part of the input) can be chosen simply as $[y_{[1]}, y_{[n]}]$, where $y_{[1]} \leq y_{[2]} \leq \cdots \leq y_{[n]}$ are the order statistics of the response variable. In experiments we set the search range even more conservatively, enlarging the sample range by 50% of length $[y_{n+1}^{\min}, y_{n+1}^{\max}] := \big[y_{[1]} - 0.25(y_{[n]} - y_{[1]}), y_{[n]} + 0.25(y_{[n]} - y_{[1]})\big]$.

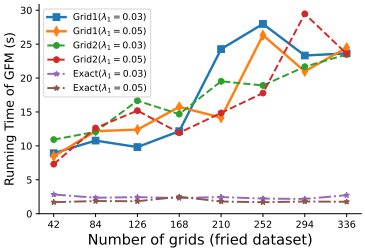

Figure 3: The training time against grid numbers.

Dataset Our experiments were conducted using real-world datasets. We employ real-world datasets from OpenML [29] and UCI repository [30] in simulations. We randomly partition the dataset into training set, testing set, and calibration set (used in `SCP`) with 70%, 10%, and 20% of the total samples. To facilitate optimization, we have standardized the entire original dataset by removing the mean and scaling to unit variance for the features, and adjusting the mean of all labels to 0.

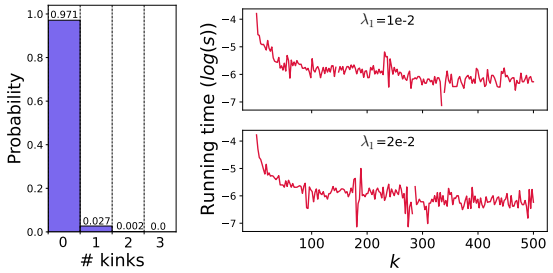

Figure 4: The histogram of kink numbers and the running time of Algorithm 1 against k, where k lives in $y_{n+1} := k \cdot z(k \in \mathbb{N})$.

Setup Our central claim is twofold, encompassing both accuracy and efficiency. We first report the average empirical coverage, average length of the prediction set in Table 2. Regarding running efficiency, we present average training time per dataset in Figure 2 while varying the scale of training set. In Figure 3, we compare the training times when different grid numbers are used in `Grid1` and `Grid2`. We further plot the histogram of kink numbers, and the running time against various $y_{n+1}(\cdot)$ in Figure 4.

Results & Analysis From Table 2 we observe that all these methods provide valid and nearly perfect coverage. The grid and exact method give similar lengths, where the slight difference is due

to the rounding between neighboring grid points. The SCP produces wider intervals due to a less efficient use of data. Given Figure 2, our exact method is much faster than the baselines, with same solid performance. We can also learn from the Figure 3 that the time of Algorithm 1 still compares favorably against the grid-search type method, even when the grid is sparse. Figure 4 found that there exists tiny number of kinks in majority runnings, which offers hope for the future expansion of our algorithm, and indicates that the choice of y_{n+1} can make a difference in efficiency, as it will determine the solving interval and the total query times of gradient.

6 Conclusion

In this work, we present a unified framework and an elaborate algorithm with statistical analysis for fast exact conformalization regarding generalized parametric estimation. We illustrate the strong and competitive performance of proposed methods in a series of benchmarks.

In future work, a potential direction is to consider scenarios where labels are multidimensional, such as multi-task learning, in which the label space $\mathcal{Y}$ would be indexed by multiple independent parameters $(z_1, \ldots, z_K)$. Under such conditions, the homotopy solution path would extend to a solution surface, and our ODE system could be reformulated as a corresponding system of partial differential equations. Additionally, compared to some previous work, Algorithm 1 offers increased speed but at the cost of higher memory requirements. It necessitates storing all training samples throughout the process for gradient queries. We believe it would be interesting to use recent advancements like the Kronecker-factored approximate method to potentially enhance the memory scalability.

Acknowledgement

I would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the presentation of this paper.

References

[1] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[2] Victor Chernozhukov, Kaspar Wüthrich, and Yinchu Zhu. Distributional conformal prediction. *Proceedings of the National Academy of Sciences*, 118(48):e2107794118, 2021.

[3] Osbert Bastani, Varun Gupta, Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Practical adversarial multivalid conformal prediction. *Advances in Neural Information Processing Systems*, 35:29362–29373, 2022.

[4] Matteo Fontana, Gianluca Zeni, and Simone Vantini. Conformal prediction: a unified review of theory and new challenges. *Bernoulli*, 29(1):1–23, 2023.

[5] Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

[6] Jonathan Alvarsson, Staffan Arvidsson McShane, Ulf Norinder, and Ola Spjuth. Predicting with confidence: using conformal prediction in drug discovery. *Journal of Pharmaceutical Sciences*, 110(1):42–49, 2021.

[7] Margaux Zaffran, Olivier Féron, Yannig Goude, Julie Josse, and Aymeric Dieuleveut. Adaptive conformal predictions for time series. In *International Conference on Machine Learning*, pages 25834–25866. PMLR, 2022.

[8] Chen Xu and Yao Xie. Conformal prediction for time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[9] Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2020.

[10] Paul Melki, Lionel Bombrun, Boubacar Diallo, Jérôme Dias, and Jean-Pierre Da Costa. Group-conditional conformal prediction via quantile regression calibration for crop and weed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 614–623, 2023.

[11] Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011.

[12] Jing Lei. Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106(4):749–764, 2019.

[13] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

[14] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[15] Geoffrey S Watson. Linear least squares regression. *The Annals of Mathematical Statistics*, pages 1679–1699, 1967.

[16] Evgeny Burnaev and Vladimir Vovk. Efficiency of conformalized ridge regression. In *Conference on Learning Theory*, pages 605–622. PMLR, 2014.

[17] Eugene Ndiaye and Ichiro Takeuchi. Computing full conformal prediction set with approximate homotopy. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Etash Kumar Guha, Eugene Ndiaye, and Xiaoming Huo. Conformalization of sparse generalized linear models. In *International Conference on Machine Learning*, pages 11871–11887. PMLR, 2023.

[19] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[20] Andrzej Ruszczynski. *Nonlinear optimization*. Princeton university press, 2011.

[21] David Avis and Komei Fukuda. A pivoting algorithm for convex hulls and vertex enumeration of arrangements and polyhedra. In *Proceedings of the seventh annual symposium on Computational geometry*, pages 98–104, 1992.

[22] James H Goodnight. A tutorial on the sweep operator. *The American Statistician*, 33(3): 149–158, 1979.

[23] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[24] Kenneth Lange, J Chambers, and W Eddy. *Numerical analysis for statisticians*, volume 2. Springer, 1999.

[25] Rasmus Bro, Sijmen De Jong, and Sijmen De Jong. A fast non-negativity-constrained least squares algorithm. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 11(5): 393–401, 1997.

[26] Xi Chen, Qihang Lin, Seyoung Kim, Jaime G Carbonell, and Eric P Xing. An efficient proximal gradient method for general structured sparse learning. *stat*, 1050, 2010.

[27] Venkata Seshadri. *The inverse Gaussian distribution: statistical theory and applications*, volume 137. Springer Science & Business Media, 2012.

[28] Parikshit Ram. On the optimality gap of warm-started hyperparameter optimization. In *International Conference on Automated Machine Learning*, pages 12–1. PMLR, 2022.

[29] Joaquin Vanschoren, Jan N Van Rijn, Bernd Bischl, and Luis Torgo. Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

[30] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[31] Fuzhen Zhang. *The Schur complement and its applications*, volume 4. Springer Science & Business Media, 2006.

[32] Dimitri Bertsekas, Angelia Nedic, and Asuman Ozdaglar. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.

[33] Michael Ulbrich. Nonsmooth newton-like methods for variational inequalities and constrained optimization problems in function spaces. *Habilitation, Technical University of Munich, Munich*, 2002.

[34] Stefan Scholtes. *Introduction to piecewise differentiable equations*. Springer, 2012.

[35] Walter Gautschi. *Numerical analysis*. Springer Science & Business Media, 2011.

[36] Jack Levine and HM Nahikian. On the construction of involutory matrices. *The American Mathematical Monthly*, 69(4):267–272, 1962.

A Proofs

A.1 Preliminaries

Definition 6 (Schur Complement). *Given a block matrix* $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$*, if* $\mathbf{A}_{22}$ *is invertible, then the Schur complement [31] of the block* $\mathbf{A}_{22}$ *of* $\mathbf{A}$ *is defined by* $\mathbf{A}/\mathbf{A}_{11} \triangleq \mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$.

Definition 7 (Local Lipschitz Continuity). *For a function* $f : \mathbb{R}^p \to \mathbb{R}$*, f is locally ℓ-Lipschitz continuous if for any bounded* $S \subseteq \mathbb{R}^p$*, there exists a positive real constant* ℓ *such that*

$$\left|f(x_1) - f(x_2)\right| \leq \ell \left\|x_1 - x_2\right\|, \ \forall x_1, x_2 \in S. \tag{16}$$

Definition 8 (Strong Convexity). *For a differentiable function* $f : \mathbb{R}^p \to \mathbb{R}$*, f is μ-strongly convex if there exists a positive real constant* μ *such that*

$$f(x_1) \geq f(x_2) + \langle \nabla f(x_2), x_1 - x_2 \rangle + \frac{\mu}{2}\left\|x_1 - x_2\right\|^2, \ \forall x_1, x_2 \in \mathbb{R}^p. \tag{17}$$

If f is twice-differentiable, μ-strong convexity is also equivalent to $\nabla^2 f(\cdot) \succeq \mu \mathbf{I}_p$.

In alignment with the Section 3.1, we focus on minimizing the surrogate function formulated as

$$\begin{aligned} \min_{\mathbf{w}\in\mathbb{R}^p} \mathcal{E}_z(\mathbf{w}) := & \underbrace{\sum_{i=1}^{n} L_i\left(y_i, \eta_{\mathbf{w}}(\mathbf{x}_i)\right) + L_{n+1}\left(y_{n+1}(z), \eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right) + \sum_{j=1}^{m} \lambda_j \Omega_j(\mathbf{w})}_{\text{model terms}} \\ & \underbrace{+ \rho \sum_{i=1}^{r} |g_i(\mathbf{w})| + \rho \sum_{j=1}^{s} \max\{0, h_j(\mathbf{w})\}}_{\text{penalty terms}}, \end{aligned} \tag{18}$$

for $\rho \in \mathbb{R}^{>0}$. We refer to the model estimation terms as $\mathbb{L}_M(\mathbf{w}|z)$ and the exact penalty terms as $\mathbb{L}_P(\mathbf{w})$. Figure 5 displays the interdependencies among key variables utilized in our analysis.

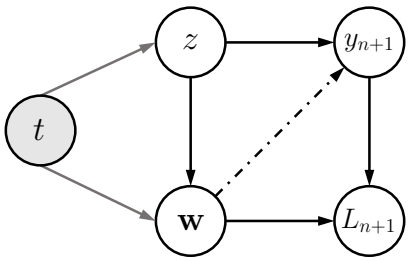

Figure 5: Diagram showing relationships between some critical variables in the paper. The grey area and arrows represent the latent reparameterization regime. The dashed arrows indicate potential relationships, that is, determined by the definition of $y_{n+1}(\cdot)$.

A.2 Proof of Theorem 2

Proof. The initial conclusion is straightforward as both $\sum_{i=1}^{r} |g_i(\cdot)|$ and $\sum_{j=1}^{s} \max\{0, h_j(\cdot)\}$ are non-negative. Regarding the second claim, consider more generally a finite family $\tilde{f}_1(\mathbf{w}), \ldots, \tilde{f}_q(\mathbf{w})$ of convex functions, and suppose a linear combination $\sum_{k=1}^{q} c_k \tilde{f}_k(\mathbf{w})$ with positive coefficients is strictly convex. It suffices to prove that any other linear combination $\sum_{k=1}^{q} b_k \tilde{f}_k(\mathbf{w})$ with positive coefficients is strictly convex. For any two distinct points $\mathbf{w}_1 \neq \mathbf{w}_2$ and any scalar $\gamma \in (0, 1)$, the following inequality holds

$$\tilde{f}_k\left[\gamma \mathbf{w}_1 + (1-\gamma)\mathbf{w}_2\right] \leq \gamma \tilde{f}_k(\mathbf{w}_1) + (1-\gamma)\tilde{f}_k(\mathbf{w}_2). \tag{19}$$

Given that $\sum_{k=1}^{q} c_k \tilde{f}_k(\mathbf{w})$ is strictly convex, a strict inequality occurs for at least one k. Hence, multiplying inequality (19) by b_k and adding yields

$$\sum_{k=1}^{q} b_k \tilde{f}_k \left[\gamma \mathbf{w}_1 + (1-\gamma)\, \mathbf{w}_2\right] < \gamma \sum_{k=1}^{q} b_k \tilde{f}_k \left(\mathbf{w}_1\right) + (1-\gamma) \sum_{k=1}^{q} b_k \tilde{f}_k \left(\mathbf{w}_2\right). \tag{20}$$

The third conclusion is derived from the principle that a convex function is coercive if and only if its restriction to each half-line is coercive, as noted in [32]. Given this result, suppose $\mathcal{E}_z(\mathbf{w})$ at ρ is coercive, but $\mathcal{E}_z(\mathbf{w})$ at ρ^* is not coercive. Then there exists a point $\mathbf{w}$, a direction $\vec{v}$, and a sequence of scalras t_n tending to ∞ such that $\mathcal{E}_z(\mathbf{w} + t_n\vec{v})$ at ρ^* is bounded above. This requires the sequence $\mathbb{L}_M\left(\mathbf{w} + t_n\vec{v}|z\right)$ and each of the sequences $|g_i(\mathbf{w} + t_n\vec{v})|$ and $\max\{0, h_j(\mathbf{w} + t_n\vec{v})\}$ to remain bounded above. But in this circumstance the sequence $\mathcal{E}_z(\mathbf{w} + t_n\vec{v})$ at ρ also remains bounded above, which opposes our assumption. The proofs of final two assertions are also straightforward and thus omitted here. □

A.3 Proof of Theorem 3

Before presenting the proof, we first introduce Lemma 2 and Lemma 3.

Lemma 2. *Let $f\colon U \to \mathbb{R}$ be a PC^r-function on the open set $U \subseteq \mathbb{R}^p$ and let C^r-functions $\{f_1, \ldots, f_k\}$ be a set of selection functions of f. Then for any $x \in U$, there exists an open neighborhood $U' \subseteq U$ of x on which f is also a continuous selection of $\left\{f_i\colon i \in I_f^e(x)\right\}$.*

Proof. Refer to a similar argument as in Proposition 2.22 of [33]. □

Lemma 3. *Let $f : U \to \mathbb{R}$ be a PC^r-function on the open set $U \subseteq \mathbb{R}^p$ and let $\{f_1, \ldots, f_k\}$ be a set of selection functions of f. Then we have*

$$\partial f(x) = \operatorname{conv}\left(\left\{\nabla f_i(x) : i \in I_f^e(x)\right\}\right), \quad \forall x \in \mathbb{R}^p. \tag{21}$$

Proof. See a similar justification as in Proposition 4.3.1 of [34]. □

With these lemmas, we are now prepared to prove Theorem 3.

Proof. According to the KKT conditions and Definition 5, the optimal point $\mathbf{w}^\star(z)$ of the function $\mathcal{E}_z\left(\mathbf{w}\right)$ is characterized by the stationarity condition

$$\begin{aligned} \mathbf{0} \in & \sum_{i=1}^{n} \partial L_i\left(y_i, \eta_{\mathbf{w}}(\mathbf{x}_i)\right) + \partial L_{n+1}\left(y_{n+1}\left(z\right), \eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right) + \sum_{j=1}^{m} \lambda_j \cdot \partial \Omega_j\left(\mathbf{w}\right) \\ & + \rho \sum_{i=1}^{r} \hat{\theta}_{g_i} \cdot \nabla g_i(\mathbf{w}) + \rho \sum_{j=1}^{s} \hat{\theta}_{h_j} \cdot \nabla h_j(\mathbf{w}), \end{aligned} \tag{22}$$

where the coefficients satisfy

$$\hat{\theta}_{g_i} \in \begin{cases} \{-1\} & g_i(\mathbf{w}) < 0 \\ [-1, 1] & g_i(\mathbf{w}) = 0 \\ \{1\} & g_i(\mathbf{w}) > 0 \end{cases}, \quad \text{and} \quad \hat{\theta}_{h_j} \in \begin{cases} \{0\} & h_j(\mathbf{w}) < 0 \\ [0, 1] & h_j(\mathbf{w}) = 0 \\ \{1\} & h_j(\mathbf{w}) > 0 \end{cases}. \tag{23}$$

Let $\mathbf{w}\left(z\right)$ be the solution of (5) indexed by the latent parameter z and $\mathbf{w}(z + \Delta z)$ the solution when the label is increased by a small amount $\Delta z > 0$. Then the difference $\Delta\mathbf{w}(z) = \mathbf{w}\left(z + \Delta z\right) - \mathbf{w}\left(z\right)$

should aim to minimize the increase in objective value. Therefore, $\Delta\mathbf{w}$ is the solution to

$$\begin{aligned}
\min_{\Delta\mathbf{w}} \quad & \mathcal{E}_{z+\Delta z}(\mathbf{w}+\Delta\mathbf{w}) - \mathcal{E}_z(\mathbf{w}) \\
&= \mathbb{L}_M(\mathbf{w}+\Delta\mathbf{w}|z(t)+\Delta z) + \mathbb{L}_P(\mathbf{w}+\Delta\mathbf{w}) - \left[\mathbb{L}_M(\mathbf{w}|z(t)) + \mathbb{L}_P(\mathbf{w})\right] \\
&= \mathbb{L}_M(\mathbf{w}|z(t)+\Delta z) + \left(\nabla\mathbb{L}_M(\mathbf{w}|z(t)+\Delta z)\right)^\top \Delta\mathbf{w} + \frac{1}{2}\Delta\mathbf{w}^\top\nabla^2\mathbb{L}_M(\mathbf{w}|z(t)+\Delta z)\,\Delta\mathbf{w} \\
&\quad + \mathbb{L}_P(\mathbf{w}+\Delta\mathbf{w}) + \sum_{\zeta=3}^{\infty}\frac{\left\langle\nabla^\zeta\mathbb{L}_M(\mathbf{w}|z(t)+\Delta z), (\Delta\mathbf{w})^\zeta\right\rangle}{\zeta!} - \mathbb{L}_M(\mathbf{w}|z(t)) - \mathbb{L}_P(\mathbf{w}) \\
&= \left[\mathbb{L}_M(\mathbf{w}|z(t)+\Delta z) - \mathbb{L}_M(\mathbf{w}|z(t))\right] + \left\langle\nabla\mathbb{L}_M(\mathbf{w}|z(t)+\Delta z), \Delta\mathbf{w}\right\rangle \\
&\quad + \frac{1}{2}\Delta\mathbf{w}^\top\nabla^2\mathbb{L}_M(\mathbf{w}|z(t)+\Delta z)\,\Delta\mathbf{w} + \left[\mathbb{L}_P(\mathbf{w}+\Delta\mathbf{w}) - \mathbb{L}_P(\mathbf{w})\right] + o\left((\Delta\mathbf{w})^3\right),
\end{aligned} \tag{24}$$

where we utilized the Taylor series expansion on $\mathbb{L}_M(\mathbf{w}+\Delta\mathbf{w}|z(t)+\Delta z)$ at the value of $\mathbf{w}$. To retain the clarity, we start from simplifying some terms in (24). First we compute the $\mathbb{L}_M(\mathbf{w}|z(t)+\Delta z) - \mathbb{L}_M(\mathbf{w}|z(t))$ as

$$\begin{aligned}
&\mathbb{L}_M(\mathbf{w}|z(t)+\Delta z) - \mathbb{L}_M(\mathbf{w}|z(t)) \\
&= \sum_{i=1}^{n} L_i\left(y_i, \eta_{\mathbf{w}}(\mathbf{x}_i)\right) + L_{n+1}\left(y_{n+1}(z(t)+\Delta z), \eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right) + \sum_{j=1}^{m}\lambda_i\Omega_i(\mathbf{w}) - \sum_{i=1}^{n} L_i\left(y_i, \eta_{\mathbf{w}}(\mathbf{x}_i)\right) \\
&\quad - L_{n+1}\left(y_{n+1}(z(t)), \eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right) - \sum_{j=1}^{m}\lambda_i\Omega_i(\mathbf{w}) \\
&= L_{n+1}\left(y_{n+1}(z(t)+\Delta z), \eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right) - L_{n+1}\left(y_{n+1}(z(t)), \eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right).
\end{aligned} \tag{25}$$

Now we revisit the definition of the following index sets determined by signs of constraint residuals

$$\begin{aligned}
\mathcal{N}_{\mathrm{E}} &= \{i : g_i(\mathbf{w}) = \mathbf{v}_i^\top\mathbf{w} - d_i < 0\}, & \mathcal{N}_{\mathrm{I}} &= \{j : h_j(\mathbf{w}) = \boldsymbol{\omega}_j^\top\mathbf{w} - e_j < 0\}, \\
\mathcal{Z}_{\mathrm{E}} &= \{i : g_i(\mathbf{w}) = \mathbf{v}_i^\top\mathbf{w} - d_i = 0\}, & \mathcal{Z}_{\mathrm{I}} &= \{j : h_j(\mathbf{w}) = \boldsymbol{\omega}_j^\top\mathbf{w} - e_j = 0\}, \\
\mathcal{P}_{\mathrm{E}} &= \{i : g_i(\mathbf{w}) = \mathbf{v}_i^\top\mathbf{w} - d_i > 0\}, & \mathcal{P}_{\mathrm{I}} &= \{j : h_j(\mathbf{w}) = \boldsymbol{\omega}_j^\top\mathbf{w} - e_j > 0\}.
\end{aligned} \tag{26}$$

Then we compute the scaled $\mathbb{L}_P(\mathbf{w}+\Delta\mathbf{w}) - \mathbb{L}_P(\mathbf{w})$ as

$$\begin{aligned}
&\frac{1}{\rho}\left[\mathbb{L}_P(\mathbf{w}+\Delta\mathbf{w}) - \mathbb{L}_P(\mathbf{w})\right] \\
&= \sum_{i\in\mathcal{P}_{\mathrm{E}}} g_i(\mathbf{w}+\Delta\mathbf{w}) + \sum_{i\in\mathcal{Z}_{\mathrm{E}}} g_i(\mathbf{w}+\Delta\mathbf{w}) - \sum_{i\in\mathcal{N}_{\mathrm{E}}} g_i(\mathbf{w}+\Delta\mathbf{w}) + \sum_{j\in\mathcal{P}_{\mathrm{I}}} h_j(\mathbf{w}+\Delta\mathbf{w}) + \sum_{j\in\mathcal{Z}_{\mathrm{I}}} h_j(\mathbf{w}+\Delta\mathbf{w}) \\
&\quad + \sum_{j\in\mathcal{N}_{\mathrm{I}}} h_j(\mathbf{w}+\Delta\mathbf{w}) - \sum_{i\in\mathcal{P}_{\mathrm{E}}} g_i(\mathbf{w}) - \sum_{i\in\mathcal{Z}_{\mathrm{E}}} g_i(\mathbf{w}) + \sum_{i\in\mathcal{N}_{\mathrm{E}}} g_i(\mathbf{w}) - \sum_{j\in\mathcal{P}_{\mathrm{I}}} h_j(\mathbf{w}) - \sum_{j\in\mathcal{Z}_{\mathrm{I}}} h_j(\mathbf{w}) - \sum_{j\in\mathcal{N}_{\mathrm{I}}} h_j(\mathbf{w}) \\
&= -\left[\sum_{i\in\mathcal{N}_{\mathrm{E}}} \mathbf{v}_i^\top(\mathbf{w}+\Delta\mathbf{w}) - d_i\right] + \left[\sum_{i\in\mathcal{P}_{\mathrm{E}}} \mathbf{v}_i^\top(\mathbf{w}+\Delta\mathbf{w}) - d_i\right] + \left[\sum_{j\in\mathcal{P}_{\mathrm{I}}} \boldsymbol{\omega}_j^\top(\mathbf{w}+\Delta\mathbf{w}) - e_j\right]
\end{aligned}$$

$$
\begin{aligned}
&+\left(\sum_{i\in\mathcal{N}_{\mathrm{E}}}\mathbf{v}_i^\top\mathbf{w}-d_i\right)+\underbrace{\sum_{i\in\mathcal{Z}_{\mathrm{E}}}g_i(\mathbf{w}+\Delta\mathbf{w})+\sum_{j\in\mathcal{Z}_{\mathrm{I}}}h_j(\mathbf{w}+\Delta\mathbf{w})-\sum_{i\in\mathcal{Z}_{\mathrm{E}}}g_i(\mathbf{w})-\sum_{j\in\mathcal{Z}_{\mathrm{I}}}h_j(\mathbf{w})}_{\text{denote as term }\mathbb{L}_{\ominus}}\\
&-\left(\sum_{i\in\mathcal{P}_{\mathrm{E}}}\mathbf{v}_i^\top\mathbf{w}-d_i\right)-\left(\sum_{j\in\mathcal{P}_{\mathrm{I}}}\boldsymbol{\omega}_j^\top\mathbf{w}-e_j\right)+\underbrace{\sum_{j\in\mathcal{N}_{\mathrm{I}}}h_j(\mathbf{w}+\Delta\mathbf{w})-\sum_{j\in\mathcal{N}_{\mathrm{I}}}h_j(\mathbf{w})}_{\text{always }0}\\
=&\left[\sum_{i\in\mathcal{N}_{\mathrm{E}}}-\left(\mathbf{v}_i^\top\mathbf{w}+\mathbf{v}_i^\top\Delta\mathbf{w}+d_i\right)+\mathbf{v}_i^\top\mathbf{w}-d_i\right]+\left[\sum_{i\in\mathcal{P}_{\mathrm{E}}}\mathbf{v}_i^\top\mathbf{w}+\mathbf{v}_i^\top\Delta\mathbf{w}-d_i-\left(\mathbf{v}_i^\top\mathbf{w}-d_i\right)\right]\\
&+\left[\sum_{j\in\mathcal{P}_{\mathrm{I}}}\boldsymbol{\omega}_j^\top\mathbf{w}+\boldsymbol{\omega}_j^\top\Delta\mathbf{w}-e_j-\left(\boldsymbol{\omega}_j^\top\mathbf{w}-e_j\right)\right]+\mathbb{L}_{\ominus}+0\\
=&-\sum_{i\in\mathcal{N}_{\mathrm{E}}}\mathbf{v}_i^\top\Delta\mathbf{w}+\sum_{i\in\mathcal{P}_{\mathrm{E}}}\mathbf{v}_i^\top\Delta\mathbf{w}+\sum_{j\in\mathcal{P}_{\mathrm{I}}}\boldsymbol{\omega}_j^\top\Delta\mathbf{w}+\mathbb{L}_{\ominus}.
\end{aligned}
\tag{27}
$$

Consequently, we obtain

$$
\begin{aligned}
\mathbb{L}_P\left(\mathbf{w}+\Delta\mathbf{w}\right)-\mathbb{L}_P\left(\mathbf{w}\right)=&\rho\cdot\underbrace{\left[-\sum_{i\in\mathcal{N}_{\mathrm{E}}}\mathbf{v}_i+\sum_{i\in\mathcal{P}_{\mathrm{E}}}\mathbf{v}_i+\sum_{j\in\mathcal{P}_{\mathrm{I}}}\boldsymbol{\omega}_j\right]^\top}_{\text{denote as }\mathbf{u}_{\bar{\mathcal{Z}}}}\cdot\Delta\mathbf{w}+\rho\mathbb{L}_{\ominus}\\
\triangleq&\,\rho\mathbf{u}_{\bar{\mathcal{Z}}}\Delta\mathbf{w}+\rho\mathbb{L}_{\ominus}.
\end{aligned}
\tag{28}
$$

Drawing on the findings from (25) and (28), we are now in a position to calculate the difference as presented in (24)

$$
\begin{aligned}
&\mathcal{E}_{z+\Delta z}(\mathbf{w}+\Delta\mathbf{w})-\mathcal{E}_z\left(\mathbf{w}\right)\\
=&L_{n+1}\left(y_{n+1}\left(z(t)+\Delta z\right),\eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right)-L_{n+1}\left(y_{n+1}\left(z(t)\right),\eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right)+\left\langle\nabla\mathbb{L}_M\left(\mathbf{w}|z(t)+\Delta z\right),\Delta\mathbf{w}\right\rangle\\
&+\rho\mathbf{u}_{\bar{\mathcal{Z}}}\Delta\mathbf{w}+\rho\mathbb{L}_{\ominus}+\frac{1}{2}\Delta\mathbf{w}^\top\nabla^2\mathbb{L}_M\left(\mathbf{w}|z(t)+\Delta z\right)\Delta\mathbf{w}+o\left((\Delta\mathbf{w})^3\right)\\
=&\sum_{i=1}^{n}\sum_{k\in I^a_{L_i}(\mathbf{w})}\hat{\theta}^k_{L_i}(\mathbf{w})\Delta\mathbf{w}^\top\nabla D^k_{L_i}\left(y_i,\eta_{\mathbf{w}}(\mathbf{x}_i)\right)+\sum_{j=1}^{m}\sum_{k\in I^a_{\Omega_j}(\mathbf{w})}\lambda_j\hat{\theta}^k_{\Omega_j}(\mathbf{w})\Delta\mathbf{w}^\top\nabla D^k_{\Omega_j}\left(\mathbf{w}\right)\\
&+\frac{1}{2}\sum_{i=1}^{n}\sum_{k\in I^a_{L_i}(\mathbf{w})}\hat{\theta}^k_{L_i}(\mathbf{w})\Delta\mathbf{w}^\top\nabla^2 D^k_{L_i}\left(y_i,\eta_{\mathbf{w}}(\mathbf{x}_i)\right)\Delta\mathbf{w}+L_{n+1}\left(y_{n+1}\left(z(t)+\Delta z\right),\eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right)\\
&+\rho\mathbf{u}_{\bar{\mathcal{Z}}}\Delta\mathbf{w}+\sum_{k\in I^a_{L_{n+1}}(\mathbf{w})}\hat{\theta}^k_{L_{n+1}}(\mathbf{w})\Delta\mathbf{w}^\top\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right),\eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right)\\
&+\frac{1}{2}\sum_{k\in I^a_{L_{n+1}}(\mathbf{w})}\hat{\theta}^k_{L_{n+1}}(\mathbf{w})\Delta\mathbf{w}^\top\nabla^2 D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right),\eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right)\Delta\mathbf{w}+\rho\mathbb{L}_{\ominus}\\
&-L_{n+1}\left(y_{n+1}\left(z(t)\right),\eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right)+\frac{1}{2}\sum_{j=1}^{m}\sum_{k\in I^a_{\Omega_j}(\mathbf{w})}\lambda_j\hat{\theta}^k_{\Omega_j}(\mathbf{w})\Delta\mathbf{w}^\top\nabla D^k_{\Omega_j}\left(\mathbf{w}\right)\Delta\mathbf{w}+o\left((\Delta\mathbf{w})^3\right).
\end{aligned}
\tag{29}
$$

To ensure that the optimal solutions adhere to the constraints and set partitions, it is necessary to establish the condition

$$
\begin{aligned}
0 \in \mathbb{L}_{\ominus}\left(\mathbf{w}^{\star}\right) &= \left[\sum_{i \in \mathcal{Z}_{\mathrm{E}}} g_i(\mathbf{w}^{\star}+\Delta\mathbf{w}^{\star}) - \sum_{i \in \mathcal{Z}_{\mathrm{E}}} g_i(\mathbf{w}^{\star})\right] + \left[\sum_{j \in \mathcal{Z}_{\mathrm{I}}} h_j(\mathbf{w}^{\star}+\Delta\mathbf{w}^{\star}) - \sum_{j \in \mathcal{Z}_{\mathrm{I}}} h_j(\mathbf{w}^{\star})\right] \\
&= \left[\sum_{i \in \mathcal{Z}_{\mathrm{E}}} \mathbf{v}_i^{\top}\mathbf{w}^{\star} + \mathbf{v}_i^{\top}\Delta\mathbf{w}^{\star} - d_i - \left(\mathbf{v}_i^{\top}\mathbf{w}^{\star} - d_i\right)\right] + \left[\sum_{j \in \mathcal{Z}_{\mathrm{I}}} \boldsymbol{\omega}_j^{\top}\mathbf{w}^{\star} + \boldsymbol{\omega}_j^{\top}\Delta\mathbf{w}^{\star} - e_j - \left(\boldsymbol{\omega}_j^{\top}\mathbf{w}^{\star} - e_j\right)\right] \\
&= \left[\sum_{i \in \mathcal{Z}_{\mathrm{E}}} \mathbf{v}_i + \sum_{j \in \mathcal{Z}_{\mathrm{I}}} \boldsymbol{\omega}_j\right]^{\top} \cdot \Delta\mathbf{w}^{\star}.
\end{aligned}
\tag{30}
$$

To the second order, $\Delta\mathbf{w}^{\star}$ is the solution to

$$
\begin{aligned}
\min_{\Delta\mathbf{w}} \quad & \mathcal{E}_{z+\Delta z}(\mathbf{w}+\Delta\mathbf{w}) - \mathcal{E}_z\left(\mathbf{w}\right) \\
\approx & \sum_{k \in I^a_{L_{n+1}}(\mathbf{w})} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}) \Delta\mathbf{w}^{\top} \nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right), \eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right) \\
& + \frac{1}{2} \sum_{k \in I^a_{L_{n+1}}(\mathbf{w})} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}) \Delta\mathbf{w}^{\top} \nabla^2 D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right), \eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right) \Delta\mathbf{w} \\
& + \Bigg\langle \underbrace{\sum_{i=1}^{n} \sum_{k \in I^a_{L_i}(\mathbf{w})} \hat{\theta}^k_{L_i}(\mathbf{w}) \nabla D^k_{L_i}\left(y_i, \eta_{\mathbf{w}}(\mathbf{x}_i)\right) + \sum_{j=1}^{m} \sum_{k \in I^a_{\Omega_j}(\mathbf{w})} \lambda_j \hat{\theta}^k_{\Omega_j}(\mathbf{w}) \nabla D^k_{\Omega_j}\left(\mathbf{w}\right)}_{\text{denote as } \mathbf{D}(\mathbf{w})}, \Delta\mathbf{w} \Bigg\rangle \\
& + \Bigg\langle \Bigg\langle \frac{\Delta\mathbf{w}}{2}, \underbrace{\sum_{i=1}^{n} \sum_{k \in I^a_{L_i}(\mathbf{w})} \hat{\theta}^k_{L_i}(\mathbf{w}) \nabla^2 D^k_{L_i}\left(y_i, \eta_{\mathbf{w}}(\mathbf{x}_i)\right) + \sum_{j=1}^{m} \sum_{k \in I^a_{\Omega_j}(\mathbf{w})} \lambda_j \hat{\theta}^k_{\Omega_j}(\mathbf{w}) \nabla^2 D^k_{\Omega_j}\left(\mathbf{w}\right)}_{\text{denote as } \mathbf{H}(\mathbf{w})} \Bigg\rangle, \Delta\mathbf{w} \Bigg\rangle \\
& + \rho \mathbf{u}_{\bar{\mathcal{Z}}} \Delta\mathbf{w} + \left(L_{n+1}\left(y_{n+1}\left(z(t)+\Delta z\right), \eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right) - L_{n+1}\left(y_{n+1}\left(z(t)\right), \eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right)\right), \\
s.t. \quad & \mathbf{v}_i^{\top}\Delta\mathbf{w} = 0, \; i \in \mathcal{Z}_{\mathrm{E}}, \\
& \boldsymbol{\omega}_j^{\top}\Delta\mathbf{w} = 0, \; j \in \mathcal{Z}_{\mathrm{I}},
\end{aligned}
\tag{31}
$$

where the equality constraints are arised from (30). This naturally results in the associated Lagrange multiplier problem

$$
\begin{bmatrix} \tilde{\mathbf{H}}(\mathbf{w}^{\star}|z,\Delta z) & \mathbf{U}_{\mathcal{Z}}^{\top} \\ \mathbf{U}_{\mathcal{Z}} & \mathbf{O}_{n_{\mathcal{Z}} \times n_{\mathcal{Z}}} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{w}^{\star} \\ \boldsymbol{\lambda}_{\mathcal{Z}} \end{bmatrix} = - \begin{bmatrix} \tilde{\mathbf{D}}(\mathbf{w}^{\star}|z,\Delta z) + \rho \mathbf{u}_{\bar{\mathcal{Z}}}^{\top} \\ \mathbf{0} \end{bmatrix},
$$

$$
\begin{aligned}
\tilde{\mathbf{H}}(\mathbf{w}^{\star}|z,\Delta z) &= \mathbf{H}(\mathbf{w}^{\star}) + \sum_{k \in I^a_{L_{n+1}}(\mathbf{w}^{\star})} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}^{\star}) \cdot \nabla^2 D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right), \eta_{\mathbf{w}^{\star}}(\mathbf{x}_{n+1})\right), \\
\tilde{\mathbf{D}}(\mathbf{w}^{\star}|z,\Delta z) &= \mathbf{D}(\mathbf{w}^{\star}) + \sum_{k \in I^a_{L_{n+1}}(\mathbf{w}^{\star})} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}^{\star}) \nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right), \eta_{\mathbf{w}^{\star}}(\mathbf{x}_{n+1})\right),
\end{aligned}
\tag{32}
$$

where $n_{\mathcal{Z}} = |\mathcal{Z}_{\mathrm{E}} \cup \mathcal{Z}_{\mathrm{I}}|$, the matrix $\mathbf{U}_{\mathcal{Z}} = \begin{bmatrix} \mathbf{v}_i^{\top} \\ \vdots \\ \boldsymbol{\omega}_j^{\top} \\ \vdots \end{bmatrix}$ comprises constant differentials for $i \in \mathcal{Z}_{\mathrm{E}}$ and

$j \in \mathcal{Z}_{\mathrm{I}}$, and the vector $\boldsymbol{\lambda}_{\mathcal{Z}}$ consists of the (introduced) Lagrange multipliers.

It is important to note that in (31), we have omitted remainder terms of order higher than 2 in Taylor series expansion, specifically $o\left((\Delta \mathbf{w})^3\right)$, for clarity. We assert that this exclusion does not lead to any theoretical loss, and the $\approx$ in (31) is effectively an equality. This might initially seem *counterintuitive*, but we will validate this assertion towards the end of this proof.

Now we denote the inverse of the matrix as[3]

$$\begin{bmatrix} \tilde{\mathbf{H}}(\mathbf{w}^\star|z,\Delta z) & \mathbf{U}_{\mathcal{Z}}^\top \\ \mathbf{U}_{\mathcal{Z}} & \mathbf{O}_{n_{\mathcal{Z}}\times n_{\mathcal{Z}}} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{P}(\mathbf{w}^\star|z,\Delta z) & \mathbf{Q}(\mathbf{w}^\star|z,\Delta z) \\ \mathbf{Q}^\top(\mathbf{w}^\star|z,\Delta z) & \mathbf{R}(\mathbf{w}^\star|z,\Delta z) \end{bmatrix}. \tag{33}$$

So the solution for the difference vector $\Delta \mathbf{w}^\star$ can be expressed via

$$\begin{aligned} \Delta \mathbf{w}^\star &= -\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\left[\tilde{\mathbf{D}}(\mathbf{w}^\star|z,\Delta z) + \rho \mathbf{u}_{\tilde{\mathcal{Z}}}^\top\right] \\ &= -\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\left[\underbrace{\mathbf{D}(\mathbf{w}^\star) + \rho \mathbf{u}_{\tilde{\mathcal{Z}}}^\top}_{\text{locally irrelevant to } \Delta z \text{ and } z(t)} + \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right), \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\right]. \end{aligned} \tag{34}$$

Applying the $\dot{z} = \xi(z(t))$, we arrive at

$$\begin{aligned} &\dot{\mathbf{w}}^\star(t) \\ &= \lim_{\Delta z\to 0} \frac{\Delta \mathbf{w}^\star(z(t))}{\Delta z}\cdot \dot{z}(t) \\ &= -\left[\lim_{\Delta z\to 0} \frac{\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\left(\mathbf{D}(\mathbf{w}^\star) + \rho \mathbf{u}_{\tilde{\mathcal{Z}}}^\top\right)}{\Delta z}\right. \\ &\qquad \left. + \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)} \lim_{\Delta z\to 0} \frac{\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\nabla D^k_{L_{n+1}}\left(y_{n+1}(z(t)+\Delta z), \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)}{\Delta z}\right]\cdot \xi(z) \\ &= -\left[\left(\partial_{\Delta z}\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\cdot\left(\mathbf{D}(\mathbf{w}^\star) + \rho \mathbf{u}_{\tilde{\mathcal{Z}}}^\top\right)\right)\Big|_{\Delta z=0}\right. \\ &\qquad + \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\lim_{\Delta z\to 0}\frac{\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z+\Delta z\right), \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)}{\Delta z} \\ &\qquad \left. + \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\cdot\lim_{\Delta z\to 0}\frac{\mathbf{P}(\mathbf{w}^\star|z,\Delta z)}{\Delta z}\cdot\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right), \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\right]\cdot \xi(z) \\ &= -\xi(z)\left(\partial_{\Delta z}\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\Big|_{\Delta z=0}\right)\left[\mathbf{D}(\mathbf{w}^\star) + \rho \mathbf{u}_{\tilde{\mathcal{Z}}}^\top\right] \\ &\quad - \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\xi(z)\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\left[\partial_{\Delta z}\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z+\Delta z\right), \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\Big|_{\Delta z=0}\right] \\ &\quad - \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\xi(z)\cdot\left(\partial_{\Delta z}\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\Big|_{\Delta z=0}\right)\cdot\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)\right), \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right) \end{aligned}$$

[3]Note that the notation here *slightly* differs from that in the main body.

$$= -\Bigg\{ \xi(z)\left(\partial_{\Delta z}\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\right)\left[\tilde{\mathbf{D}}(\mathbf{w}^\star|z,\Delta z)+\rho\mathbf{u}_{\mathcal{Z}}^\top\right] + \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\xi(z)\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\cdot\left[\partial_{\Delta z}\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\right]\Bigg\}_{\Delta z=0}. \tag{35}$$

Utilizing (8) and recognizing that $\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^\top=\mathbf{0}$, we can proceed to simplify the results

$$\begin{aligned}
\dot{\mathbf{w}}^\star(t) =& -\Bigg\{\xi(z)\left(\partial_{\Delta z}\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\right)\cdot\Bigg[\sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\\
&-\sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)-\rho\mathbf{U}_{\mathcal{Z}}^\top\boldsymbol{r}_{\mathcal{Z}}\Bigg]\\
&+\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\cdot\Bigg[\sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\xi(z)\left[\partial_{\Delta z}\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\right]\\
&-\sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\xi(z)\left[\partial_{\Delta z}\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\right]-\mathbf{0}\Bigg]\Bigg\}_{\Delta z=0}\\
=& -\Bigg\{\sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\xi(z)\cdot\Bigg[\left(\partial_{\Delta z}\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\right)\Big[\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\\
&-\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\Big]+\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\Big[\partial_{\Delta z}\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\\
&-\partial_{\Delta z}\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\Big]\Bigg]\Bigg\}_{\Delta z=0}\\
=& -\sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\xi(z)\Bigg\{\left(\partial_{\Delta z}\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\right)\mathbf{D}^{\varkappa}(\mathbf{w}^\star|z,\Delta z)+\mathbf{P}(\mathbf{w}^\star|z,\Delta z)\cdot\\
&\left(\partial_{\Delta z}\mathbf{D}^{\varkappa}(\mathbf{w}^\star|z,\Delta z)\right)\Bigg\}_{\Delta z=0}\\
=& -\sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\xi(z)\mathbf{P}(\mathbf{w}^\star|z,0)\cdot\left[\partial_z\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\right],
\end{aligned} \tag{36}$$

where $\mathbf{D}^{\varkappa}(\mathbf{w}^\star|z,\Delta z)$ is given by the $(n+1)$-th gradient difference

$$\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z(t)+\Delta z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)-\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right).$$

Next, we examine the remainder term $o\left((\Delta\mathbf{w})^3\right)$, which can be explicitly expressed as

$$o\left((\Delta \mathbf{w})^3\right) = \frac{1}{6} \cdot \left\langle \nabla^3 \mathbb{L}_M \left(\mathbf{w} | z(t) + \Delta z\right), (\Delta \mathbf{w})^3 \right\rangle + \sum_{\zeta=4}^{\infty} \frac{\left\langle \nabla^\zeta \mathbb{L}_M \left(\mathbf{w} | z(t) + \Delta z\right), (\Delta \mathbf{w})^\zeta \right\rangle}{\zeta!}. \tag{37}$$

When we incorporate this into our analysis, the left-hand side of the first equation in (32) will include the additional term

$$\left\langle \left[\nabla \tilde{\mathbf{H}},\ \nabla^2 \tilde{\mathbf{H}},\ \ldots,\ \nabla^{(\zeta-2)} \tilde{\mathbf{H}}\right]^\top,\ \left[\frac{1}{3!}\left(\Delta \mathbf{w}^\star\right)^2,\ \frac{1}{4!}\left(\Delta \mathbf{w}^\star\right)^3,\ \ldots,\ \frac{\left(\Delta \mathbf{w}^\star\right)^{\zeta-1}}{(\zeta-1)!}\right]^\top \right\rangle. \tag{38}$$

Following our previous analysis and recalling that

$$\Delta \mathbf{w}^\star(z) := \mathbf{w}^\star\left(z + \Delta z\right) - \mathbf{w}^\star\left(z\right),$$

we obtain

$$\begin{aligned}
&\lim_{\Delta z \to 0} \frac{\left\langle \left[\nabla \tilde{\mathbf{H}},\ \nabla^2 \tilde{\mathbf{H}},\ \ldots,\ \nabla^{(\zeta-2)} \tilde{\mathbf{H}}\right]^\top,\ \left[\frac{1}{3!}\left(\Delta \mathbf{w}^\star\right)^2,\ \frac{1}{4!}\left(\Delta \mathbf{w}^\star\right)^3,\ \ldots,\ \frac{\left(\Delta \mathbf{w}^\star\right)^{\zeta-1}}{(\zeta-1)!}\right]^\top \right\rangle}{\Delta z} \\
&= \lim_{\Delta z \to 0} \left\langle \frac{\Delta \mathbf{w}^\star}{\Delta z},\ \left\langle \left[\nabla \tilde{\mathbf{H}},\ \nabla^2 \tilde{\mathbf{H}},\ \ldots,\ \nabla^{(\zeta-2)} \tilde{\mathbf{H}}\right]^\top,\ \left[\frac{1}{6}\Delta \mathbf{w}^\star,\ \frac{1}{24}\left(\Delta \mathbf{w}^\star\right)^2,\ \ldots,\ \frac{\left(\Delta \mathbf{w}^\star\right)^{\zeta-2}}{(\zeta-1)!}\right]^\top \right\rangle \right\rangle \\
&= \lim_{\Delta z \to 0} \left\langle \frac{\Delta \mathbf{w}^\star}{\Delta z},\ \left\langle \left[\nabla \tilde{\mathbf{H}},\ \nabla^2 \tilde{\mathbf{H}},\ \ldots,\ \nabla^{(\zeta-2)} \tilde{\mathbf{H}}\right]^\top,\right.\right. \\
&\qquad \left.\left. \left[\frac{1}{6}\left(\mathbf{w}^\star\left(z+\Delta z\right) - \mathbf{w}^\star\left(z\right)\right),\ \frac{1}{24}\left(\mathbf{w}^\star\left(z+\Delta z\right) - \mathbf{w}^\star\left(z\right)\right), \ldots,\ \frac{\left(\mathbf{w}^\star\left(z+\Delta z\right) - \mathbf{w}^\star\left(z\right)\right)^{\zeta-2}}{(\zeta-1)!}\right]^\top \right\rangle \right\rangle \\
&= \lim_{\Delta z \to 0} \left\langle \frac{\Delta \mathbf{w}^\star}{\Delta z}, \left\langle \left[\nabla \tilde{\mathbf{H}}, \nabla^2 \tilde{\mathbf{H}},\ \ldots, \nabla^{(\zeta-2)} \tilde{\mathbf{H}}\right]^\top, \left[\frac{\mathbf{w}^\star(z) - \mathbf{w}^\star(z)}{6},\ \ldots, \frac{\left(\mathbf{w}^\star(z) - \mathbf{w}^\star(z)\right)^{\zeta-2}}{(\zeta-1)!}\right]^\top \right\rangle \right\rangle \\
&= \lim_{\Delta z \to 0} \left\langle \underbrace{\frac{\Delta \mathbf{w}^\star}{\Delta z}}_{\neq \pm\infty},\ \left\langle \underbrace{\left[\nabla \tilde{\mathbf{H}},\ \nabla^2 \tilde{\mathbf{H}},\ \ldots,\ \nabla^{(\zeta-2)} \tilde{\mathbf{H}}\right]^\top}_{\neq \pm\infty},\ \mathbf{0} \right\rangle \right\rangle \\
&= \lim_{\Delta z \to 0} \left\langle \frac{\Delta \mathbf{w}^\star}{\Delta z},\ \mathbf{0} \right\rangle = 0,
\end{aligned} \tag{39}$$

which implies that the remainder term will have no impact on the final derived system of differential equations. Thus, (13) furnishes an exact solution towards the whole conformal path $\{\mathbf{w}^\star(z) : z = z(t), 0 \leq t \leq \mathbb{T}\}$, thereby concluding our proof. □

A.4 Proof of Theorem 4

Proof. The existence and uniqueness of the solution in (13) can be referenced in the proof of Theorem 5.3.1 in [35], which derives its basis from the Picard–Lindelöf theorem. We now revisit the following expression as

$$\min_{\mathbf{w} \in \mathbb{R}^p} \mathcal{E}_z\left(\mathbf{w}\right) \triangleq \mathbb{L}_M\left(\mathbf{w} | z\right) + \mathbb{L}_P\left(\mathbf{w}\right). \tag{40}$$

Considering the optimal solution, where $\mathbf{w}^\star$ represents the minimizer of $\mathcal{E}_z\left(\mathbf{w}\right)$, the following inequality holds

$$\mathcal{E}_{z+\Delta z}\left(\mathbf{w}^\star\left(z + \Delta z\right)\right) \leq \mathcal{E}_{z+\Delta z}\left(\mathbf{w}^\star\left(z\right)\right). \tag{41}$$

The $\mathcal{E}_{z+\Delta z}\left(\mathbf{w}^\star\left(z\right)\right)$ can be rewritten as $\mathcal{E}_z\left(\mathbf{w}^\star\left(z\right)\right) + \tilde{o}(\Delta z)$, wherein $\lim_{\Delta z \to 0} \tilde{o}(\Delta z) = 0$.[4] By adding

[4]The decomposition is logical when $\lim_{\Delta z \to 0} \mathcal{E}_z\left(\mathbf{w}^\star\left(z\right)\right) + \tilde{o}(\Delta z) = \mathcal{E}_z\left(\mathbf{w}^\star\left(z\right)\right) = \lim_{\Delta z \to 0} \mathcal{E}_{z+\Delta z}\left(\mathbf{w}^\star\left(z\right)\right)$, implying that the $(n+1)$-th loss $L_{n+1}\left(\cdot\right)$ is continuous with respect to z. This continuity is easily verified upon the fullfillment of our mild assumptions.

$\mathcal{E}_z\left(\mathbf{w}^\star\left(z+\Delta z\right)\right)$ to both sides of the (41), it turns to

$$\mathcal{E}_z\left(\mathbf{w}^\star\left(z+\Delta z\right)\right)-\mathcal{E}_z\left(\mathbf{w}^\star\left(z\right)\right)\leq \underbrace{\mathcal{E}_z\left(\mathbf{w}^\star\left(z+\Delta z\right)\right)-\mathcal{E}_{z+\Delta z}\left(\mathbf{w}^\star\left(z+\Delta z\right)\right)}_{\text{denote as term } \mathcal{E}^{\ominus}(\mathbf{w}^\star)}+\tilde{o}(\Delta z). \tag{42}$$

Let

$$\hat{z}:=\arg\max_{z}\ \mathcal{E}^{\ominus}\left(\mathbf{w}^\star\left(z\right)\right), \tag{43}$$

we have

$$\begin{aligned}\mathcal{E}^{\ominus}\left(\mathbf{w}^\star\left(z\right)\right)\leq\ & \mathcal{E}_z\left(\mathbf{w}^\star\left(\hat{z}\right)\right)-\mathcal{E}_{z+\Delta z}\left(\mathbf{w}^\star\left(\hat{z}\right)\right)\\ &\leq \mathbb{L}_M\left(\mathbf{w}^\star|\hat{z}\right)+\mathbb{L}_P\left(\mathbf{w}^\star\right)-\left[\mathbb{L}_M\left(\mathbf{w}^\star|\hat{z}+\Delta z\right)+\mathbb{L}_P\left(\mathbf{w}^\star\right)\right]\\ &=\mathbb{L}_M\left(\mathbf{w}^\star|\hat{z}\right)-\mathbb{L}_M\left(\mathbf{w}^\star|\hat{z}+\Delta z\right)\\ &\leq L_{n+1}\left(y_{n+1}\left(\hat{z}\right),\eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right)-L_{n+1}\left(y_{n+1}\left(\hat{z}+\Delta z\right),\eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right).\end{aligned} \tag{44}$$

In accordance with the assumption made, it is imperative for $y_{n+1}(z)$ to maintain continuity concerning z, or

$$\lim_{\Delta z\to 0}\left(y_{n+1}\left(\hat{z}\right)-y_{n+1}\left(\hat{z}+\Delta z\right)\right)=0. \tag{45}$$

Consequently, we can deduce that

$$\begin{aligned}&\lim_{\Delta z\to 0}\mathcal{E}_z\left(\mathbf{w}^\star\left(z+\Delta z\right)\right)-\mathcal{E}_z\left(\mathbf{w}^\star\left(z\right)\right)\\ \leq\ &\lim_{\Delta z\to 0}\mathcal{E}_z\left(\mathbf{w}^\star\left(z+\Delta z\right)\right)-\mathcal{E}_{z+\Delta z}\left(\mathbf{w}^\star\left(z+\Delta z\right)\right)+\tilde{o}(\Delta z)\\ \leq\ &\lim_{\Delta z\to 0}\left[L_{n+1}\left(y_{n+1}\left(\hat{z}\right),\eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right)-L_{n+1}\left(y_{n+1}\left(\hat{z}+\Delta z\right),\eta_{\mathbf{w}}(\mathbf{x}_{n+1})\right)\right]+\tilde{o}(\Delta z)\\ =\ &0.\end{aligned} \tag{46}$$

As evident from (40), it is clear that $\mathcal{E}_z\left(\mathbf{w}^\star\left(z+\Delta z\right)\right)-\mathcal{E}_z\left(\mathbf{w}^\star\left(z\right)\right)\geq 0$. Thus we get

$$\lim_{\Delta z\to 0}\mathcal{E}_z\left(\mathbf{w}^\star\left(z+\Delta z\right)\right)-\mathcal{E}_z\left(\mathbf{w}^\star\left(z\right)\right)=0. \tag{47}$$

Due to the uniqueness of the ODE solution, the (47) further implies

$$\lim_{\Delta z\to 0}\mathbf{w}^\star(z+\Delta z)=\mathbf{w}^\star(z). \tag{48}$$

Following the definition of continuity, we can conclude that $\mathbf{w}^\star$ is continuous at z.

Additionally, in the context of coefficient continuity, and given the assumption of linear independence, the unique solutions for $\hat{\theta}_{g_i}(z)$ and $\hat{\theta}_{h_j}(z)$ can be obtained by applying the stationarity condition (8) given the solution vector $\mathbf{w}^\star(z)$. Consequently, the continuity of $\hat{\theta}_{g_i}(z)$ and $\hat{\theta}_{h_j}(z)$ is inherited from the continuity of $\mathbf{w}^\star(z)$. □

A.5 Proof of Theorem 5

Proof. The (8) implies

$$\begin{aligned}&\mathbf{D}(\mathbf{w}^\star)+\sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)\\ =&-\rho\sum_{i=1}^{r}\hat{\theta}_{g_i}\nabla g_i(\mathbf{w})-\rho\sum_{j=1}^{s}\hat{\theta}_{h_j}\nabla h_j(\mathbf{w})\\ =&-\rho\mathbf{u}^\top_{\mathcal{Z}}-\rho\mathbf{U}^\top_{\mathcal{Z}}\boldsymbol{r}_{\mathcal{Z}}.\end{aligned} \tag{49}$$

Multiplying both sides by $\mathbf{Q}(\mathbf{w}^\star|z,\Delta z)$ yields

$$\mathbf{Q}(\mathbf{w}^\star|z,\Delta z)\mathbf{D}(\mathbf{w}^\star)+\sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\mathbf{Q}(\mathbf{w}^\star|z,\Delta z)\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right),\eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right)$$

$$
\begin{aligned}
&= -\mathbf{Q}(\mathbf{w}^\star|z,\Delta z)\rho\mathbf{u}_{\bar{\mathcal{Z}}}^\top - \rho\left(\mathbf{Q}(\mathbf{w}^\star|z,\Delta z)\mathbf{U}_{\bar{\mathcal{Z}}}^\top\right)\boldsymbol{r}_{\mathcal{Z}} \\
&= -\mathbf{Q}(\mathbf{w}^\star|z,\Delta z)\rho\mathbf{u}_{\bar{\mathcal{Z}}}^\top - \rho\boldsymbol{r}_{\mathcal{Z}}.
\end{aligned} \tag{50}
$$

We can now proceed to solve for $\boldsymbol{r}_{\mathcal{Z}}$ from (50) as follows

$$
\begin{aligned}
\boldsymbol{r}_{\mathcal{Z}} &= -\mathbf{Q}(\mathbf{w}^\star|z,\Delta z=0)\left[\frac{1}{\rho}\mathbf{D}(\mathbf{w}^\star) + \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}^\star)}\frac{1}{\rho}\hat{\theta}^k_{L_{n+1}}(\mathbf{w}^\star)\nabla D^k_{L_{n+1}}\left(y_{n+1}(z), \eta_{\mathbf{w}^\star}(\mathbf{x}_{n+1})\right) + \mathbf{u}_{\bar{\mathcal{Z}}}^\top\right] \\
&= -\mathbf{Q}(\mathbf{w}^\star|z,0)\left[\frac{1}{\rho}\tilde{\mathbf{D}}(\mathbf{w}^\star|z,\Delta z=0) + \mathbf{u}_{\bar{\mathcal{Z}}}^\top\right],
\end{aligned} \tag{51}
$$

which is equivalent to (14). □

A.6 Proof of Theorem 6

Before presenting our formal proof, we first introduce Lemma 4 and Lemma 5.

Lemma 4. *Let $\{f_i\}_{i=1}^K$ be a finite collection of functions, where each $f_i : \mathbb{R}^p \to \mathbb{R}$ is ℓ_i-Lipschitz continuous. We have the linear combination of these functions*

$$
g(x) := \sum_{i=1}^K \theta_i f_i(x),\ \theta_i > 0,\ 1 \le i \le K, \tag{52}
$$

is also Lipschitz continuous with a Lipschitz constant $\sum_{i=1}^K \theta_i\ell_i$.

Proof. Consider any two points $x_1, x_2 \in \mathbb{R}^p$, we have

$$
\begin{aligned}
\left|g(x_1) - g(x_2)\right| &= \left|\sum_{i=1}^K \theta_i f_i(x_1) - \sum_{i=1}^K \theta_i f_i(x_2)\right| \\
&= \left|\sum_{i=1}^K \theta_i \cdot (f_i(x_1) - f_i(x_2))\right|.
\end{aligned} \tag{53}
$$

Using the ℓ_i-Lipschitz continuity of each f_i, we get

$$
\begin{aligned}
\left|g(x_1) - g(x_2)\right| &\le \left|\sum_{i=1}^K \theta_i \cdot \ell_i \left\|x_1 - x_2\right\|\right| \\
&= \sum_{i=1}^K \theta_i\ell_i \cdot \left\|x_1 - x_2\right\|.
\end{aligned} \tag{54}
$$

Thus, by Definition 7, $g(\cdot)$ is Lipschitz continuous with Lipschitz constant $\ell_g = \sum_{i=1}^K \theta_i\ell_i$. □

Lemma 5. *Let $\{f_i\}_{i=1}^K$ be a finite collection of functions, where each $f_i : \mathbb{R}^p \to \mathbb{R}$ is μ_i-strongly convex. We have the linear combination of these functions*

$$
g(x) := \sum_{i=1}^K \theta_i f_i(x),\ \theta_i > 0,\ 1 \le i \le K, \tag{55}
$$

is also strongly convex with a strong convexity constant $\sum_{i=1}^K \theta_i\mu_i$.

Proof. To prove this, consider any two points $x_1, x_2 \in \mathbb{R}^p$. By the property of strong convexity for each f_i, we have

$$f_i(x_1) \geq f_i(x_2) + \langle \nabla f_i(x_2), x_1 - x_2 \rangle + \frac{\mu_i}{2} \left\| x_1 - x_2 \right\|^2, \ 1 \leq i \leq K. \tag{56}$$

Multiplying each inequality by positive θ_i and summing over i, we can obtain

$$\begin{aligned} \sum_{i=1}^{K} \theta_i f_i(x_1) &\geq \sum_{i=1}^{K} \theta_i \left[f_i(x_2) + \langle \nabla f_i(x_2), x_1 - x_2 \rangle + \frac{\mu_i}{2} \left\| x_1 - x_2 \right\|^2 \right] \\ &= \sum_{i=1}^{K} \theta_i f_i(x_2) + \left\langle \nabla \sum_{i=1}^{K} \theta_i f_i(x_2), x_1 - x_2 \right\rangle + \sum_{i=1}^{K} \frac{\theta_i \mu_i}{2} \left\| x_1 - x_2 \right\|^2. \end{aligned} \tag{57}$$

By definition of $g(\cdot)$, (57) simplifies to

$$g(x_1) \geq g(x_2) + \langle \nabla g(x_2), x_1 - x_2 \rangle + \frac{1}{2} \sum_{i=1}^{K} \theta_i \mu_i \cdot \left\| x_1 - x_2 \right\|^2, \ \forall x_1, x_2 \in \mathbb{R}^p. \tag{58}$$

Thus, $g(\cdot)$ is strongly convex with convexity constant $\mu_g = \sum_{i=1}^{K} \theta_i \mu_i$ by Definition 8. □

With these lemmas, we are now ready to prove Theorem 6.

Proof. Considering any two valid solutions $\mathbf{w}_1^\star$ and $\mathbf{w}_2^\star$, the function difference of $\Upsilon(\cdot, \cdot)$ can be expanded as

$$\begin{aligned} &\left\| \Upsilon(\mathbf{w}_1^\star, z) - \Upsilon(\mathbf{w}_2^\star, z) \right\| \\ =& \Bigg\| \Upsilon(\mathbf{w}_1^\star, z) + \sum_{k \in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star) \xi(z) \mathbf{P}(\mathbf{w}_1^\star | z) \cdot \left[\partial_z \nabla D^k_{L_{n+1}} \left(y_{n+1}(z), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1}) \right) \right] \\ &\quad - \sum_{k \in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star) \xi(z) \mathbf{P}(\mathbf{w}_1^\star | z) \cdot \left[\partial_z \nabla D^k_{L_{n+1}} \left(y_{n+1}(z), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1}) \right) \right] - \Upsilon(\mathbf{w}_2^\star, z) \Bigg\| \\ \leq& \underbrace{\left\| \Upsilon(\mathbf{w}_1^\star, z) + \sum_{k \in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star) \xi(z) \mathbf{P}(\mathbf{w}_1^\star | z) \cdot \left[\partial_z \nabla D^k_{L_{n+1}} \left(y_{n+1}(z), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1}) \right) \right] \right\|}_{\text{denote as term } \clubsuit} \\ &+ \underbrace{\left\| - \sum_{k \in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star) \xi(z) \mathbf{P}(\mathbf{w}_1^\star | z) \cdot \left[\partial_z \nabla D^k_{L_{n+1}} \left(y_{n+1}(z), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1}) \right) \right] - \Upsilon(\mathbf{w}_2^\star, z) \right\|}_{\text{denote as term } \spadesuit}, \end{aligned} \tag{59}$$

where the last step involves the application of triangle inequality. We then bound the term ♣ as

$$\begin{aligned} \clubsuit =& \Bigg\| \xi(z) \cdot \mathbf{P}(\mathbf{w}_1^\star | z) \cdot \Bigg[- \sum_{k \in I^a_{L_{n+1}}(\mathbf{w}_1^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_1^\star) \left[\partial_z \nabla D^k_{L_{n+1}} \left(y_{n+1}(z), \eta_{\mathbf{w}_1^\star}(\mathbf{x}_{n+1}) \right) \right] \\ &\quad + \sum_{k \in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star) \left[\partial_z \nabla D^k_{L_{n+1}} \left(y_{n+1}(z), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1}) \right) \right] \Bigg] \Bigg\| \end{aligned}$$

$$
\begin{aligned}
&\leq \left\|\xi(z)\right\| \cdot \left\|\mathbf{P}(\mathbf{w}_1^\star|z)\right\| \cdot \Bigg\| \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}_1^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_1^\star) \left[\partial_z \nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right), \eta_{\mathbf{w}_1^\star}(\mathbf{x}_{n+1})\right)\right] \\
&\quad - \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star) \left[\partial_z \nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1})\right)\right] \Bigg\| \\
&= \left\|\xi(z)\right\| \cdot \left\|\mathbf{P}(\mathbf{w}_1^\star|z)\right\| \cdot \Bigg\| \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star) \Bigg[\partial_z \Bigg[\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right), \eta_{\mathbf{w}_1^\star}(\mathbf{x}_{n+1})\right)\Bigg] \\
&\quad - \left[\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1})\right)\right]\Bigg] \Bigg\| \\
&\leq \bar{C} \cdot \left\|\mathbf{P}(\mathbf{w}_1^\star|z)\right\| \cdot \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star) \Bigg\| \partial_z \left[\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right), \eta_{\mathbf{w}_1^\star}(\mathbf{x}_{n+1})\right)\right] \\
&\quad - \left[\nabla D^k_{L_{n+1}}\left(y_{n+1}\left(z\right), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1})\right)\right] \Bigg\| \\
&\leq \bar{C} \cdot \left\|\mathbf{P}(\mathbf{w}_1^\star|z)\right\| \cdot \ell \left\|\mathbf{w}_1^\star - \mathbf{w}_2^\star\right\|,
\end{aligned} \tag{60}
$$

where the last two derivations utilize the Lemma 4 and fact that $\sum_{k\in I^a_{L_i}(\mathbf{w}^\star)} \hat{\theta}^k_{L_i}(\mathbf{w}^\star) - 1 = 0$. Referring to (70), we can bound the $\left\|\mathbf{P}(\mathbf{w}_1^\star|z)\right\|$ as

$$
\begin{aligned}
\left\|\mathbf{P}(\mathbf{w}_1^\star|z)\right\| &= \left\|\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z) - \tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1} \mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right\| \\
&\leq \left\|\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right\| + \left\|\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1} \mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right\|.
\end{aligned} \tag{61}
$$

By employing Lemma 5 and recognizing that $\sum_{k\in I^a_{\Omega_j}(\mathbf{w}^\star)} \hat{\theta}^k_{\Omega_j}(\mathbf{w}^\star) - 1 = 0$, $\sum_{k\in I^a_{L_i}(\mathbf{w}^\star)} \hat{\theta}^k_{L_i}(\mathbf{w}^\star) - 1 = 0$, we arrive at

$$
\begin{aligned}
\left\|\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right\| &\leq \frac{1}{\sum_{i=1}^{n+1} \sum_{k\in I^a_{L_i}(\mathbf{w}_1^\star)} \hat{\theta}^k_{L_i}(\mathbf{w}_1^\star)\cdot\mu + \sum_{j=1}^{m} \sum_{k\in I^a_{\Omega_j}(\mathbf{w}_1^\star)} \lambda_j \hat{\theta}^k_{\Omega_j}(\mathbf{w}_1^\star)\cdot\sigma} \\
&\leq \frac{1}{(n+1)\mu + \sum_{j=1}^{m}\lambda_j\sigma}.
\end{aligned} \tag{62}
$$

Then we need to bound $\left\|\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1} \mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right\|$ in (61), shown as

$$
\begin{aligned}
&\left\|\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1} \mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right\| \\
&\leq \underbrace{\left\|\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1} \mathbf{U}_{\mathcal{Z}}\right\|}_{\text{denote as term } \|\blacklozenge\|} \cdot \left\|\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right\|.
\end{aligned} \tag{63}
$$

One might note that

$$\begin{aligned}
\blacklozenge \cdot \blacklozenge &= \tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1}\mathbf{U}_{\mathcal{Z}} \cdot \tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1}\mathbf{U}_{\mathcal{Z}} \\
&= \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right]^\top \cdot \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1} \cdot \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right] \cdot \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1}\mathbf{U}_{\mathcal{Z}} \\
&= \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right]^\top \cdot \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1} \cdot \mathbf{I} \cdot \left[\mathbf{U}_{\mathcal{Z}}^\top\right]^\top \\
&= \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1} \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right] \\
&= \mathbf{I},
\end{aligned} \tag{64}$$

where this observation is based on the property that $\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top = \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^\top$. This implies the $\blacklozenge$ is an involutory matrix, as indicated in [36]. By the prior studies in [36], the eigenvalues of $\blacklozenge$ are therefore $\{1, -1\}$. Thus, we conclude that

$$\begin{aligned}
&\left\|\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1}\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right\| \\
&\leq \|\blacklozenge\| \cdot \left\|\tilde{\mathbf{H}}^{-1}(\mathbf{w}_1^\star|z)\right\| \\
&\leq 1 \cdot \frac{1}{(n+1)\mu + \sum_{j=1}^{m}\lambda_j\sigma},
\end{aligned} \tag{65}$$

and

$$\begin{aligned}
\clubsuit &\leq \bar{C} \cdot \left\|\mathbf{P}(\mathbf{w}_1^\star|z)\right\| \cdot \ell \left\|\mathbf{w}_1^\star - \mathbf{w}_2^\star\right\| \\
&\leq 2\bar{C}\ell \left[(n+1)\mu + \sum_{j=1}^{m}\lambda_j\sigma\right]^{-1} \cdot \left\|\mathbf{w}_1^\star - \mathbf{w}_2^\star\right\|.
\end{aligned} \tag{66}$$

To establish a boundary for the term $\spadesuit$, we once again invoke Lemma 4 and the fact that $\sum_{k\in I^a_{L_i}(\mathbf{w}^\star)} \hat{\theta}^k_{L_i}(\mathbf{w}^\star) - 1 = 0$. Now we can deduce the following

$$\begin{aligned}
\spadesuit &= \left\| \left[\mathbf{P}(\mathbf{w}_1^\star|z) - \mathbf{P}(\mathbf{w}_2^\star|z)\right] \cdot \left[- \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star)\xi(z) \left[\partial_z \nabla D^k_{L_{n+1}}\left(y_{n+1}(z), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1})\right)\right]\right] \right\| \\
&\leq \left\|\mathbf{P}(\mathbf{w}_1^\star|z) - \mathbf{P}(\mathbf{w}_2^\star|z)\right\| \cdot \left\|\xi(z)\right\| \cdot \left\| \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star) \left[\partial_z \nabla D^k_{L_{n+1}}\left(y_{n+1}(z), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1})\right)\right]\right\| \\
&\leq \ell \left\|\mathbf{w}_1^\star - \mathbf{w}_2^\star\right\| \cdot \bar{C} \cdot \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star) \left\|\partial_z \left[\nabla D^k_{L_{n+1}}\left(y_{n+1}(z), \eta_{\mathbf{w}_2^\star}(\mathbf{x}_{n+1})\right)\right]\right\| \\
&\leq \bar{C}\ell \left\|\mathbf{w}_1^\star - \mathbf{w}_2^\star\right\| \cdot \sum_{k\in I^a_{L_{n+1}}(\mathbf{w}_2^\star)} \hat{\theta}^k_{L_{n+1}}(\mathbf{w}_2^\star)\ell \\
&= \bar{C}\ell^2 \left\|\mathbf{w}_1^\star - \mathbf{w}_2^\star\right\|.
\end{aligned} \tag{67}$$

In conclusion, the difference of $\Upsilon(\cdot,\cdot)$ can be bounded by

$$\left\|\Upsilon(\mathbf{w}_1^\star, z) - \Upsilon(\mathbf{w}_2^\star, z)\right\| \leq \clubsuit + \spadesuit, \tag{68}$$

where

$$\begin{aligned}\clubsuit + \spadesuit &\leq 2\bar{C}\ell \left[(n+1)\mu + \sum_{j=1}^{m} \lambda_j \sigma\right]^{-1} \cdot \left\|\mathbf{w}_1^\star - \mathbf{w}_2^\star\right\| + \bar{C}\ell^2 \cdot \left\|\mathbf{w}_1^\star - \mathbf{w}_2^\star\right\| \\ &= \left[\bar{C}\ell^2 + \frac{2\bar{C}\ell}{(n+1)\mu + \sum_{j=1}^{m} \lambda_j \sigma}\right] \left\|\mathbf{w}_1^\star - \mathbf{w}_2^\star\right\|.\end{aligned} \tag{69}$$

By employing the Definition 7, we complete the proof. □

A.7 Proof of sweeping

Prior to presenting our detailed analysis, we introduce Lemma 6.

Lemma 6. *Suppose U is an invertible matrix with shape $p \times p$, and V is a $n \times n$ matrix. The inverse matrix of $\begin{bmatrix} U & Z^T \\ Z & V \end{bmatrix}$ equals to $\begin{bmatrix} U^{-1} + U^{-1}Z^T\widetilde{H}^{-1}ZU^{-1} & -U^{-1}Z^T\widetilde{H}^{-1} \\ -\widetilde{H}^{-1}ZU^{-1} & \widetilde{H}^{-1} \end{bmatrix}$, if $\widetilde{H} = V - ZU^{-1}Z^T$ is an invertible Schur complement.*

Proof. The multiplication of the aforementioned matrices, irrespective of the order of operation, yields a matrix with dimension conforming to the shape $\begin{bmatrix} \mathbf{I}_p & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_n \end{bmatrix}$. □

Subsequently, we proceed with our analysis to elucidate why the sweeping technique can yield the desired shape for specific key matrices.

Proof. Utilizing (33) and Lemma 6, we can explicitly express the matrix components of $\mathbf{P}$, $\mathbf{Q}$ and $\mathbf{R}$ as follows

$$\begin{aligned}\mathbf{P}(\mathbf{w}^\star|z,\Delta z) &= \tilde{\mathbf{H}}^{-1}(\mathbf{w}^\star|z,\Delta z) - \tilde{\mathbf{H}}^{-1}(\mathbf{w}^\star|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^\top \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^\star|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1} \mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^\star|z,\Delta z) \\ \mathbf{Q}(\mathbf{w}^\star|z,\Delta z) &= \tilde{\mathbf{H}}^{-1}(\mathbf{w}^\star|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^\top \left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^\star|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1} \\ \mathbf{R}(\mathbf{w}^\star|z,\Delta z) &= -\left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^\star|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^\top\right]^{-1}.\end{aligned} \tag{70}$$

Recalling that performing a sweeping operation on the k-th diagonal entry, where $A_{kk} \neq 0$, of matrix $\mathbf{A}$ results in a matrix $\widehat{\mathbf{A}}$ with entries

$$\begin{aligned}\hat{A}_{kk} &= -\frac{1}{A_{kk}}, & \hat{A}_{ik} &= \frac{A_{ik}}{A_{kk}},\ i \neq k, \\ \hat{A}_{kj} &= \frac{A_{kj}}{A_{kk}},\ j \neq k, & \hat{A}_{ij} &= A_{ij} - \frac{A_{ik}A_{kj}}{A_{kk}},\ \ i,j \neq k.\end{aligned} \tag{71}$$

Meanwhile, the inverse sweep operation transforms $\mathbf{A}$ into $\check{\mathbf{A}}$ with entries given by

$$\begin{aligned}\check{A}_{kk} &= -\frac{1}{A_{kk}}, & \check{A}_{ik} &= -\frac{A_{ik}}{A_{kk}}, \\ \check{A}_{kj} &= -\frac{A_{kj}}{A_{kk}},\ j \neq k, & \check{A}_{ij} &= A_{ij} - \frac{A_{ik}A_{kj}}{A_{kk}},\ \ i,j \neq k.\end{aligned} \tag{72}$$

Based on (71) and (72), we can conclude that

$$\left[\begin{array}{c|c} \tilde{\mathbf{H}}(\mathbf{w}^\star|z,\Delta z) & * \\ \hline \mathbf{U}_{\mathcal{Z}} & \mathbf{O}_{n_{\mathcal{Z}} \times n_{\mathcal{Z}}} \end{array}\right] \xrightarrow{\text{sweep}} \left[\begin{array}{c|c} -\tilde{\mathbf{H}}^{-1}(\mathbf{w}^\star|z,\Delta z) & * \\ \hline \mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^\star|z,\Delta z) & -\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^\star|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^\top \end{array}\right], \tag{73}$$

and then

$$\left[\begin{array}{c|c} \mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^{\star}|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^{\top} & \mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^{\star}|z,\Delta z) \\ \hline * & \tilde{\mathbf{H}}^{-1}(\mathbf{w}^{\star}|z,\Delta z) \end{array}\right] \xrightarrow{\text{sweep}}$$

$$\left[\begin{array}{c|c} -\left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^{\star}|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^{\top}\right]^{-1} & \tilde{\mathbf{H}}^{-1}(\mathbf{w}^{\star}|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^{\top}\left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^{\star}|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^{\top}\right]^{-1} \\ \hline * & \tilde{\mathbf{H}}^{-1}(\mathbf{w}^{\star}|z,\Delta z)-\tilde{\mathbf{H}}^{-1}(\mathbf{w}^{\star}|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^{\top}\left[\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^{\star}|z,\Delta z)\mathbf{U}_{\mathcal{Z}}^{\top}\right]^{-1}\mathbf{U}_{\mathcal{Z}}\tilde{\mathbf{H}}^{-1}(\mathbf{w}^{\star}|z,\Delta z) \end{array}\right]$$

$$\xrightarrow{\text{rearrange}} \quad \left[\begin{array}{c|c} \mathbf{R}(\mathbf{w}^{\star}|z,\Delta z) & \mathbf{Q}^{\top}(\mathbf{w}^{\star}|z,\Delta z) \\ \hline * & \mathbf{P}(\mathbf{w}^{\star}|z,\Delta z) \end{array}\right]. \tag{74}$$

This aligns with our approach outlined in Section 4.1. □

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: See Section 1 and Section 3.

   Guidelines:

   - The answer NA 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 NA 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: See Section 6.

   Guidelines:

   - The answer NA 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: See Section 2, Section 3 and Appendix A.

Guidelines:

- The answer NA 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 contribution is primarily a new statistical framework. See Section 4 and Section 5.

   Guidelines:

   - The answer NA 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: This paper does not release new assets.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 5.

   Guidelines:

   - The answer NA 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: See Section 5.

   Guidelines:

   - The answer NA 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: See Section 5.

   Guidelines:

   - The answer NA 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: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA 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: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA 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: Licenses are available referring to the provided links.

    Guidelines:

    - The answer NA 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: [NA] .

Justification: This paper does not release new assets.

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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: [NA]

Justification: [NA]

Guidelines:

- The answer NA 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.