Coupled Learning Enabled Stochastic Programming with Endogenous Uncertainty*

Junyi Liu Guangyu Li Suvrajeet Sen

Original Aug 2020

Abstract

Predictive analytics, empowered by machine learning, is usually followed by decision-making problems in prescriptive analytics. We extend the above sequential prediction-optimization paradigm to a coupled scheme such that the prediction model can guide the decision problem to produce coordinated decisions yielding higher levels of performance. Specifically, for stochastic programming (SP) models with latently decision-dependent uncertainty, we develop a coupled learning enabled optimization (CLEO) algorithm in which the learning step of predicting the latent dependency and the optimization step of computing a candidate decision are conducted interactively. The CLEO algorithm automatically balances the exploration and exploitation via the trust region method with active sampling. Under certain assumptions, we show that the sequence of solutions provided by CLEO converges to a directional stationary point of the original SP problem with probability 1. In addition, we present preliminary experimental results which demonstrate the computational potential of this data-driven approach.

1 Introduction

Beginning with George Dantzig's well-known formulation in the mid 1950's, stochastic programming (SP) has provided a class of optimization models in which uncertainties are modeled by probability distributions. Since then, due to the enormous amount of decision-making problems in the presence of uncertainty in practice, the SP field has flourished with theoretical and algorithmic advances [5, 38, 41]. Without knowing the exact probability distribution of the uncertainty, SP methods utilize discretization with scenarios of the uncertainty. Current studies of SP methods pay little attention to the generation process underlying scenarios. Instead SP methods simply assume that scenarios are generated a priori. This is reasonable when the uncertainty is decision-independent and scenarios can be generated from a historical data set or simulated from prediction models. For two-stage and multi-stage stochastic optimization problems with the decision *independent* uncertainty, there is an abundant literature [22, 25] on scenario reduction and scenario generation to ensure that the scenario tree is not too large, and yet representative of the uncertainty being modeled. Several works by Grossman and the coauthors such as [20] consider a scenario tree based method for a type of endogenous uncertainty such that the optimization decision variables influence the time of information discovery for discrete uncertainty. In the present paper, another type of endogenous uncertainty is considered, meaning that the probability distribution of the random variable is dependent on the decision variable.

*All authors are affliated with University of Southern California. Email: junyiliu@usc.edu, guangyul@usc.edu, s.sen@usc.edu.

Specifically, we focus on a task of solving an SP equipped with an *uncharacterized* uncertainty which has a dependency on the continuous decision variable:

$$\begin{array}{ll} \text{minimize} & f(x) \triangleq c(x) + \mathbf{E}_{\widetilde{\omega}|x}[\, h(x,\widetilde{\omega}) + g(x,\widetilde{\omega}\,)\,] \\ \text{subject to} & x \in X \subseteq \mathbb{R}^p \end{array} \tag{1}$$

where $\widetilde{\omega} \in \mathbb{R}^m$ is a continuous random vector defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P}(x))$, with Ω being the sample space, $\mathcal{F}$ being the σ-algebra generated by subsets of Ω and $\mathbb{P}(x)$ being a probability measure parameterized by x defined on $\mathcal{F}$. This type of uncertainty is also called endogenous uncertainty. At the outset, one needs to be sure that the conditional expectation in (1) is well defined. For a general treatment of this issue, the reader may consult [41, Section 2.3], in particular Theorem 7.37.

We are interested in the setting when the random variable is a regression model of the decision variable, i.e., $\widetilde{\omega} = m(x,\widetilde{\varepsilon}) = \psi(x) + \widetilde{\varepsilon}$ where: $\psi(\cdot) : Z \subseteq \mathbb{R}^p \to \mathbb{R}^m$, Z is an open convex set containing the feasible set X and $\widetilde{\varepsilon}$ is a random variable independent of x. With the regression model, the problem (1) has a standard SP formulation as follows.

$$\underset{x\in X\subseteq\mathbb{R}^p}{\text{minimize}} \quad f(x) \,=\, c(x) + \mathbf{E}_{\widetilde{\varepsilon}}[\, h(x,\psi(x)+\widetilde{\varepsilon}\,)\,] + \mathbf{E}_{\widetilde{\varepsilon}}[\, g(x,\psi(x)+\widetilde{\varepsilon}\,)\,]. \tag{2}$$

However, standard SP approaches cannot be applied, since both the regression function ψ and the distribution of the random variable $\widetilde{\varepsilon}$ are unknown, and moreover the realizations of $\widetilde{\varepsilon}$ cannot be observed. In order to solve such SP problems effectively, data pairs of decisions and uncertainties should be collected together to *learn* the relationship of uncertainty with the decision variable.

A practical method is "Predict-then-Optimize" (PO) ([18]). In this sequential paradigm, with a given set of data pairs $\{(x^i,\omega^i)\}$, one first predicts the probability distribution of the random variable parameterized by the decision variable, and then solves the SP problem composed with the predicted probability distribution. However, there are several technical issues associated with the PO scheme: a) in most practical cases, the true probability distribution of uncertainty may not be approximated well by parametric models; b) it may be expensive to acquire data pairs in the entire decision space, hence nonparametric prediction models may not approximate the true probability distribution of uncertainty well; c) it is computationally challenging to solve the coupled nonconvex and nonsmooth composite optimization problem in the PO scheme and the obtained solution is not guaranteed when the prediction errors are significant.

In fact, the interdependence between predictions and decisions are at the center of the trade-off between predictive accuracy and mathematical optimization of decision-making. So a question that arises is: *without any parametric assumption on the latent dependency, what class of composite prediction-optimization structure could be amenable to algorithmic optimization for solving the SP with latently endogenous uncertainty?* We address this question by coupling learning models with iterative optimization algorithms with several major contributions summarized as below.

1. Without making any parametric assumption of the probability distribution of decision-dependent uncertainty, we develop an algorithm called "Coupled Learning Enabled Optimization" (CLEO) and the sequence of solutions produced by CLEO is shown to converge to a stationary solution of the original nonlinear, non-convex and non-smooth constrained SP problem (1) with probability 1.

2. Instead of acquiring data pairs of decisions and uncertainties in the entire decision space beforehand, the CLEO algorithm acquires data pairs in a sequence of trust regions (TR) in order to iteratively learn the local dependency within local regions. Numerical experiments show that the total number of data pairs required in CLEO could be significantly lower on average than the data size required in the PO schemes.

3. In the CLEO scheme, on one hand, local linear regression model is locally comparable to the nonparametric prediction models and locally better than parametric prediction models. On the other hand, the composite optimization subproblem with a local linear regression model is a convex problem in the trust region, thus a simpler optimization problem to solve. This connection between learning and optimization in CLEO automatically balances the trade-off between prediction accuracy and optimization complexity.

4. Central points and radius of small regions for local learning models are automatically determined by the ratio of improved performance at solutions obtained in iterative TR optimization steps. This connection between learning and optimization balances the trade-off between exploring new solution region and exploiting the current solution region. We note that this is analogous to the exploitation-exploration trade-off in reinforcement learning, but this type of trade-off is novel in stochastic programming with endogenous uncertainty.

1.1 Problem setting

We make several assumptions about the program (1) so that we focus on the treatment of latent dependency.

(A1) The regression function $\psi(\cdot) : Z \subseteq \mathbb{R}^p \to \mathbb{R}^m$ is a twice-differentiable function, $\widetilde{\varepsilon}$ is a random vector independent of x, with a compact sample space. It has zero mean and finite variance Σ.

(A2) The function $h(x, \omega) = \max_{j \in \mathcal{J}} h_j(x, \omega)$. The set $\mathcal{J}$ is a finite index set and for each $j \in \mathcal{J}$, $h_j(\cdot, \cdot) : Z_j \times \Omega_j \to \mathbb{R}$ is a convex and C^1 smooth function where $Z_j \subseteq \mathbb{R}^p$ is an open convex set containing X and $\Omega_j \subseteq \mathbb{R}^m$ is an open convex set containing Ω.

(A3) The function $c(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is C^1 smooth on X and the function $g(\cdot, \cdot)$ is C^1 smooth on $X \times \Omega$. The functions c, g, h and ψ have the Lipschitz continuity modulus L_1, and c, g, $\{h_j\}_{j \in \mathcal{J}}$ and ψ have Lipschitz gradient modulus L_2.

(A4) The feasible set $X \subseteq \mathbb{R}^p$ is a deterministic, convex and compact set

We provide some motivation for the above assumptions. In assumption (A1), we assume that the random variable $\widetilde{\omega}$ is generated from a homoscedastic regression model which means that the noise $\widetilde{\varepsilon}$ has a finite variance independently of x. Notice that we do not have any parametric assumption on the regression function ψ. Given the widespread use of regression models, this assumption can easily leverage modern tools of statistical learning. Assumptions (A2) and (A3) characterize properties of the objective function $f(x)$. In particular, $c(\cdot)$ is a smooth but nonconvex function independent of the randomness. There are two functions associated with randomness: $g(\cdot, \cdot)$ is a smooth but potentially nonconvex function and $h(\cdot, \cdot)$ is a convex and nonsmooth function due to the structure of finite pointwise maximization. Such functions arise in some applications such as two-stage stochastic linear programs (SLP). Assumption (A4) imposes a deterministic and convex feasible solution set so that we focus on the treatment of the latent decision-dependency.

This type of latently decision-dependent uncertainty appears: (a) in marketing where x may denote an advertising plan and ω may denote sales [39]; (b) in revenue management where x may denote a booking policy and ω may denote demands [13]; (c) in disaster preplanning where x may denote a preparatory plan and ω may denote the survival of links after a disaster [35] and so on. In order to make the connection with standard statistical/machine learning terminology, we treat x as the predictor, and ω as the response. To seek an approximate solution of (1), we will estimate the dependency of the response on the predictor using an entirely data-driven approach.

Though problem (2) shares some relationships with composite optimization problems (such as [15]) or the finite sum in regularized empirical risk minimization problems (such as [36]), the most significant difference of (2) from the previous problems is that the objective function $f(x)$ has an implicitly composite structure due to the unknown endogenous uncertainty. The composition of a nonsmooth function h with the unknown nonconvex function ψ integrated with the randomness $\widetilde{\varepsilon}$ leads to the significant challenges in solving an implicitly-composite, nonsmooth and nonconvex SP problem (2).

To motivate the class of applications of SP with implicit endogenous uncertainty, we discuss a joint production and pricing example. The demand is a random variable which is implicitly affected by the price. Numerical performance of CLEO on this joint production and pricing problem is presented in Section 4. Besides, a joint production, shipment and pricing problem can be formulated as a two-stage SLP with endogenous uncertainty as well and we refer the interested readers to [3] for details.

Example: Joint production and pricing problem

Suppose we have a set of K products to be launched. For each product i, we need to decide its price p_i and the amount q_i to produce at the cost of c_{1i} per unit. The demand d_i of each product i is not only affected by its own price but also by the price of other products. We use bold vectors $\mathbf{p}, \mathbf{q}, \mathbf{c}_1$ and $\mathbf{d}$ to represent the price, production units, production cost, and demand for K products. Suppose that the demand $\mathbf{d}$ is a regression model of the price $\mathbf{p}$ in the form that $\mathbf{d} = \psi(\mathbf{p}) + \widetilde{\varepsilon}$ where $\widetilde{\varepsilon}$ is a random variable with bounded variance and is independent of $\mathbf{p}$ and $\mathbf{q}$. Due to the variability of the random demand, we have the option of last-minute production at the cost of $\mathbf{c}_2 > \mathbf{c}_1$ per unit to satisfy the demand. Moreover, we suffer the penalty cost of $\mathbf{c}_3$ per unit if the production exceeds the demand. The goal is to decide the price and production quantity to minimize the expected total cost. This problem is modeled as a stochastic programming problem as follows. It is straightforward to notice that this problem satisfies assumptions (A1)-(A4).

$$\underset{\mathbf{q},\mathbf{p}\geq\mathbf{0}}{\text{minimize}} \quad \mathbf{c}_1^\top \mathbf{q} + \mathbf{E}_{\mathbf{d}|p}[\, -\mathbf{p}^\top \mathbf{d} + \mathbf{c}_2^\top \max\{\mathbf{d}-\mathbf{q}, \mathbf{0}\} + \mathbf{c}_3^\top \max\{\mathbf{q}-\mathbf{d}, \mathbf{0}\}\,] \tag{3}$$

In practice, besides the price, the demand could depend on features such as seasonality and promotion. Actually, these features can be categorized into two classes, controllable features (e.g., price, promotion) and environmental features (e.g., seasonality). When only controllable features are considered in modeling the uncertainty, the CLEO scheme could be applied for such situations. When both price and seasonality features are considered, the goal is to seek the best plan of production and pricing for a specified season. Mathematically, the SP model aims to minimize $\mathbf{E}_{\mathbf{d}|(\mathbf{p},\bar{z})}\,[\, f(\mathbf{p},\mathbf{q},\mathbf{d})\,]$, where $\bar{z}$ represents a specific seasonality and $f(\mathbf{p},\mathbf{q},\mathbf{d}) \triangleq \mathbf{c}_1^\top \mathbf{q} - \mathbf{p}^\top \mathbf{d} + \mathbf{c}_2^\top \max\{\mathbf{d}-\mathbf{q}, \mathbf{0}\} + \mathbf{c}_3^\top \max\{\mathbf{q}-\mathbf{d}, \mathbf{0}\}$. If we can obtain data pairs of $(\mathbf{p}, \mathbf{d})$ for the specified season $\bar{z}$, then the CLEO scheme is suitable. If we can only obtain the data pairs of $(\mathbf{p}, z, \mathbf{d})$, then the CLEO scheme might need further adaptations. However, the study of the SP problem accounting for both controllable and environmental features is beyond the scope of the present paper.

1.2 Literature review

Decision-dependent uncertainty is common in operations management (OM) models. In revenue management, Cooper et al. [13] analyzed that if the dependency of demand and price is ignored, this modeling error could result in the systematic deterioration when the decisions are made based on the prediction using observed data over time. In dynamic pricing, there is a growing literature [1, 10] which combines demand learning and pricing to control the regret. We refer to [14] for a comprehensive review of literature in dynamic pricing. Most of the work in revenue management assumes known or parametric demand functions while there are a few works [4, 8] addressing the nonparametric demand models. In newsvendor-type problems, Lee et al. [28] utilize the quantile structure to design an iterative decision process which is combined with a forecasting model. Liyanage and Shanthikumar [31] proposes operational statistics by integrating parameter estimation and optimization to obtain a better solution compared with the traditional approach for the newsvendor inventory control problem with ambiguous demand. In these works, the integration of learning and optimization relies on special structures of these operations management models.

Connections with Stochastic Programming In stochastic programming, Dupačová [16] categorized endogenous uncertainty into two subclasses, where the decision can either affect the probability distribution of the uncertainty or affect the time when the information of the uncertainty is revealed. In the current literature, the decision-dependent information revelation has received the most attention (see [20, 23]). As for decision-dependent probability distribution, the literature is quite sparse since nonconvexity can easily creep in. Dupačová [16] addresses two classes of problems where the decision either affects parameters of the parametric probability distribution or affects the choice of a probability distribution among a finite set of probability distributions. Hellemo et al. [23] add another class of problems where the probabilities are distorted by decision variables. Noyan et al. [34] study the distributionally robust optimization model with decision-dependent uncertainty in risk measures such as CVaR. Current studies of endogenous uncertainty in SP either take the dependency of uncertainty on the decision variable being provided a priori, or have parametric assumptions of dependency.

In contrast to the above modeling-oriented viewpoints in SP, we do not make any parametric assumptions on the endogenous uncertainty. We integrate the local prediction model of the unknown endogenous uncertainty models with the decision-making using a trust-region type algorithm in this paper. There are also sampling approaches which mimic trust-region type algorithms via adaptive proximal parameter choices for stochastic linear programming (e.g., [24]) and more recently [30]. In the former paper, an iterative choice of the proximal parameter is akin to the iterative choice of a trust region radius. In the present paper, the need for model-fitting within endogenous uncertainty models in this paper, makes the trust region approach a more natural choice.

Predict-then-Optimize(PO) To accommodate the latent dependency, we take the decision variable as the predictor and the uncertainty as the response. PO schemes of this nature are very common in practice. For instance, in revenue management, a common approach is to first approximate the demand curve by a linear function of price (see [10]) and then plug the approximate linear demand curve into pricing problems. Actually, techniques used for SP with endogenous uncertainty in most of the literature can be seen as PO approaches. In such a paradigm, a choice among various prediction models affects the approximate optimization problem and thus affects the quality of the derived decision. In the case of exogenous uncertainty, Elmachtoub and Grigas

[18] develop "smart" predict-then-optimize approach which modifies the loss function of prediction models by accounting for the objective value of a special class of linear programming problems. Since PO approaches deal with prediction and optimization separately, one needs a novel way of balancing the trade-off between predictive accuracy and optimization complexity.

Derivative-free methods Instead of estimating the latent dependency, one could use derivative-free methods [12, 27] without knowing function structures. In [19], a zeroth-order method is applied to compute a critical point for the stochastic programming problem with known decision-dependent uncertainty under a smoothness condition. Another applicable method is trust region (TR) methods based on derivative-free models. We refer to Conn et al. [11] for comprehensive study of trust region methods. A classical trust region method iteratively solves a subproblem using the second-order approximation in the trust region centered at the proposed solution. As the algorithm proceeds, the trust region radius is either contracted or expanded depending on the ratio of actual-to-predicted reduction on the objective value. Without using derivative and the second-order information of the objective function, we can construct a random trust region model using either interpolation or regression models over decision variables. Recent studies [2, 6, 9, 21, 26] develop various trust region methods with random models for convergence to a first-order and second-order stationary solution of smooth and unconstrained optimization problems. In addition, a type of derivative-free approach was developed by Bertsimas et al. [3] which approximates the conditional expectation of a random function given imperfect observations using machine learning (ML) methods including LOESS, regression trees and random forests. In the case of decision-dependent uncertainty, the approach in [3] overcomes non-convexity by discretizing the space of continuous decision variables and aggregating the optimal value of subproblems for all discretized decisions. Such discretization of the decision space introduces exponential growth in the size of decision-making problem.

The CLEO algorithm, which is developed in the present paper, is based on a type of derivative-free trust region method for solving SP with latently decision-dependent uncertainty. Specifically, the CLEO algorithm is comprised of an iterative scheme consisting of learning steps and optimization steps. In a learning step, we predict the local latent dependency using local linear regression (LLR) centered at the current iterate. In the optimization step we seek a candidate solution of a trust region subproblem composed with the random LLR model. The difference of CLEO with the current literature on derivative-free trust region methods such as [9] is that the proposed TR-based method is adapted for nonsmooth, latently composite, constrained stochastic programming problems. The other work relevant to our approach is [26] in which a derivative-free optimization method is developed by utilizing the trust region framework with regression models from data. The setting considered here is different from the above reference due to a latent composite structure in SP problem caused by the decision-dependent uncertainty.

Local linear regression is an old approach dating back to late 19th century, which approximates the mean of dependent variables locally by a member in a parametric function class (see [32]). Hence, we can estimate a much wider class of regression surfaces than parametric regression models. The bandwidth of local regression models is traditionally chosen to be a constant or to contain a fixed number of points as a means of balancing the bias and variance of estimation. In its more recent incarnation, local regression includes weights using a kernel based on a given bandwidth. In CLEO, we use the uniform kernel in local linear regression, then the bandwidth represents the radius of the region for fitting a linear regression model. Moreover, we integrate learning with optimization by allowing the bandwidth of LLR model to follow the trust region radius rule so that it is automatically adjusted to the ratio of actual-to-predicted reduction. Such interaction provides a bridge between

the accuracy of local learning models and the optimality of optimization subproblems. Though a local regression model with the second or higher order functions could provide higher accuracy, it would result in a more challenging nonconvex composite subproblem. In an effort to control the complexity of any iteration, we use simple local linear regression without much loss of accuracy. To the best of our knowledge, the CLEO algorithm is the first to present a provably convergent algorithm for coupling learning and stochastic optimization to control the overall complexity in seeking a near-optimal solution for SP with decision-dependent uncertainty.

2 The CLEO algorithm

2.1 Notations

In the following analysis, we use $\mathbf{1}_{p\times q}$ to denote the matrix of size $p \times q$ with all one entries, $\mathbf{1}_p$ to denote the vector of size p with all one entries, $\mathbf{0}_{p\times q}$ to denote the matrix of size $p \times q$ with all zero entries and $\mathbb{I}_p$ denote the identity matrix of size p. For a vector $v \in \mathbb{R}^m$, $\|v\|$ denotes the Euclidean norm of a vector. For a matrix $M \in \mathbb{R}^{m\times n}$, $\|M\|$ denotes the matrix norm induced by the Euclidean vector norm. We follow the classical $O(\cdot)$ and $o(\cdot)$ representations for asymptotic behavior. For a smooth function $g(\cdot,\cdot) : X_1 \times X_2 \to \mathbb{R}$, we use $\nabla_1 g(x_1, x_2)$ to denote the gradient taken with respect to the first variable at (x_1, x_2) and $\nabla_2\, g(x_1, x_2)$ to denote the gradient taken with respect to the second variable at (x_1, x_2). For a convex function $h(\cdot,\cdot) : X_1 \times X_2 \to \mathbb{R}$, we use $\partial_1\, h(x_1, x_2)$ nad $\partial_2\, h(x_1, x_2)$ to denote the subdifferentials taken with respect to the first variable and the second variable at (x_1, x_2) respectively. When we address the probability distributions of random variables, we use $\mathcal{MN}_{m,n}(M, U, V)$ to denote the matrix normal distribution and $\mathcal{U}(S)$ to denote the uniform distribution on a set S.

Because problem (2) is potentially a nonconvex optimization problem, we limit our scope to seeking a directional stationary point which relieves the computational burden of searching for a global or local optimum. With the fact that a composition of convex and differentiable function $h(x, \psi(x)+\widetilde{\varepsilon})$ is Clarke regular and locally Lipschitz continuous, we can compute the directional derivative of f by [15] and Theorem 7.44 in [41]. For any $x \in X$ and any $d \in \mathcal{T}(x; X)$ where $\mathcal{T}(x; X)$ defines the tangent cone of the set X at x, the directional derivative of f at x for the direction d is computed as follows.

$$f'(x;\, d) = \nabla c(x)^\top d + \mathbf{E}_{\widetilde{\varepsilon}}[\, h'((x, \psi(x) + \widetilde{\varepsilon});\, (d, \nabla\psi(x)^\top d)) + g'((x, \psi(x) + \widetilde{\varepsilon});\, (d, \nabla\psi(x)^\top d))\,] \quad (4)$$

where

$$\begin{aligned} h'((x, \psi(x) + \widetilde{\varepsilon});\, (d, \nabla\psi(x)^\top d)) &= \sup\{q_1^\top d + q_2^\top \nabla\psi(x)\, d : q_1 \in \partial_1 h(x, \psi(x) + \widetilde{\varepsilon}), q_2 \in \partial_2 h(x, \psi(x) + \widetilde{\varepsilon})\} \\ g'((x, \psi(x) + \widetilde{\varepsilon});\, (d, \nabla\psi(x)^\top d)) &= \nabla_1 g\,(x, \psi(x) + \widetilde{\varepsilon})^\top d + \nabla_2 g\,(x, \psi(x) + \widetilde{\varepsilon})^\top \nabla\psi(x)^\top d \end{aligned}$$

Accordingly, the directional stationarity of (2) is defined as follows which can be found in Chapter 8 in [37] by Rockafellar and Wets. We refer to [29] for a recent review of stationarity in nonconvex problems.

Definition 1. *We say that $\bar{x}$ is a directional stationary point of the problem (2) if*

$$f'(\bar{x};\, x - \bar{x}\,) \geq 0, \quad \textit{for any } x \in X,$$

or equivalently, $\chi(x) = 0$ with

$$\chi(x) \triangleq |\, \min\{f'(x;\, d) : x + d \in X,\, \|\, d\,\| \leq 1\,\}| \quad (5)$$

Since $\psi(\cdot)$ and the probability distribution of $\widetilde{\varepsilon}$ are not available a priori, our strategy couples the estimation of a regression model with the optimization in (2). In the CLEO algorithm, we iteratively conduct a learning step with LLR and an optimization step using the TR method. In the rest of this section, we touch upon LLR models and the trust region method in turn.

2.2 LLR models

We assume that we are able to obtain data pairs of the decisions and uncertainties in local regions as the algorithm proceeds. In practice, this can be achieved by simulations from the complex model of the random variable $\widetilde{\omega}$. In learning steps, we consider a hypothesis class of affine functions, i.e.,

$$\mathcal{H} \triangleq \{\, \phi(\cdot) : \mathbb{R}^p \to \mathbb{R}^m \mid \phi(x) = (B^1)^\top x + (B^0)^\top, \text{ and } B^1 \in \mathbb{R}^{p\times m}, B^0 \in \mathbb{R}^{1\times m} \,\}.$$

At the kth iteration, given the current point $\widehat{x}^k$ and the current trust region radius δ_k, suppose we draw the data set $T_k = \{(x^i, \omega^i)\}$ of size N_k with the set of points $\{x^i\}_{i=1}^{N_k}$ uniformly generated in the local region $\in \mathcal{B}(\widehat{x}^k, \delta_k) \cap X$. Among parametric functions in the hypothesis class $\mathcal{H}$, we seek the one which minimizes the sum of square errors on the data set T_k. One may use a kernel to weight data points according to their distance to the current point $\widehat{x}^k$ or use weighted least squares to deal with heteroscedasticity when the variance of the noise in regression is dependent on x. In the present paper, under the homoscedasticity assumption, the simplest variant is to use a uniform kernel for local linear regression. Then, the estimation of parameters $\widehat{B}^{k,1}$, $\widehat{B}^{k,0}$ and residuals $\{e^{k,i}\}$ are constructed as follows.

$$\{\widehat{B}^{k,0}, \widehat{B}^{k,1}\} \in \operatorname*{argmin}_{B^0,B^1} \sum_{(x^i,\omega^i)\in T_k} \left\| \omega^i - (B^1)^\top x^i - (B^0)^\top \right\|^2, \tag{6}$$

$$e^{\,k,i} \leftarrow \omega^i - (\widehat{B}^{k,1})^\top x^i - (\widehat{B}^{k,0})^\top, \quad \text{for } i = 1, 2, \ldots, N_k \tag{7}$$

Let $\widehat{\varepsilon}^k$ denote a random variable with the empirical probability distribution of $\{e^{k,i}\}_{i=1}^{N_k}$. The LLR model m_k is constructed as follows.

$$m_k(\widehat{x}^k + s,\, \widehat{\varepsilon}^k) \triangleq (\widehat{B}^{k,1})^\top(\widehat{x}^k + s) + (\widehat{B}^{k,0})^\top + \widehat{\varepsilon}^k. \tag{8}$$

2.3 TR models

The TR model f_k is then constructed with $\mathbf{E}_{\widehat{\varepsilon}^k}$ denoting the expectation with respect to the empirical probability distribution of $\{e^{k,i}\}_{i=1}^{N_k}$.

$$f_k(\widehat{x}^k + s) \triangleq c(\widehat{x}^k + s) + \mathbf{E}_{\widehat{\varepsilon}^k}\,[\, h(\widehat{x}^k + s, m_k(\widehat{x}^k + s, \widehat{\varepsilon}^k)) + g(\widehat{x}^k + s, m_k(\widehat{x}^k + s, \widehat{\varepsilon}^k))\,]. \tag{9}$$

From an intuitive perspective, the LLR model m_k is a first-order probabilistically accurate approximation of latent dependency with enough number of samples. By composite structure, the trust region model f_k is a probabilistically accurate approximation to the original objective function. Such a criterion on accuracy will be formally discussed in the convergence analysis in section 3.1. A trust region subproblem is formulated as follows.

$$\begin{aligned} (\mathbf{P_k}) \qquad & \underset{s}{\text{minimize}} \quad f_k(\widehat{x}^k + s) \\ & \text{subject to} \quad \widehat{x}^k + s \in X, \;\; \| s \| \le \delta_k \end{aligned} \tag{10}$$

Note that under assumptions $(A2)$ and $(A3)$, f_k is the sum of two smooth functions and a convex nonsmooth function. If the smooth part is convex, the trust region subproblem (10) is a convex problem and we could find a global optimum by using any efficient numerical algorithms for convex programs, such as the stochastic approximation algorithm [33], the stochastic decomposition algorithm [40, 30] and others. If the smooth part is nonconvex, we can compute a suitable point towards the steepest descent direction that satisfies some "sufficient" decrease condition. Specifically, following the construction in Chapter 11 of [11], for any given point $\widehat{x}^k \in X$, we say s^k is a suitable step of $(\mathbf{P}_k)$ if $\widehat{x}^k + s \in X$, $\|s^k\| \leq \delta_k$; moreover, there exists a positive constant $\kappa_{dcp} \in (0,1)$ such that,

$$f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k) \;\geq\; \kappa_{dcp}\, \chi_k(\widehat{x}^k)\, \min\{\,\delta_k, 1\}\,, \quad \text{for any } \bar{x} \in \mathcal{B}(\widehat{x}^k, \epsilon) \cap X, \tag{11}$$

where

$$\chi_k(x) \triangleq |\underset{\substack{x+d \in X \\ \|d\| \leq 1}}{\text{minimize}}\, f_k{}'(x;\, d)\,|. \tag{12}$$

Theorem 12.2.2 in [11] implies that the "sufficient" decrease condition (11) can be satisfied by a step to the boundary of the steepest descent direction. The method to derive such a point is beyond the scope of our paper; instead we refer the interested readers to section 12.2 in Conn et al. [11].

Because of modeling error, one needs to verify whether the derived suitable step s_k is a descent step. Trust region methods use the ratio of actual-to-predicted reduction as a criterion. Due to the unknown decision-dependency, actual function values can not be computed, but can be estimated using local regression models. Specifically, to estimate function values $f(\widehat{x}^k)$ and $f(\widehat{x}^k + s^k)$, we first generate two independent data sets $S_k = \{(x^{j,1}, \omega^{j,1})\}$ and $S_{k+1/2} = \{(x^{j,2}, \omega^{j,2})\}$ with two sets of decision points $\{x^{j,1}\}$ and $\{x^{j,2}\}$ uniformly generated in neighborhoods $\mathcal{B}(\widehat{x}^k, \|s_k\|/2) \cap X$ and $\mathcal{B}(\widehat{x}^k + s^k, \|s_k\|/2) \cap X$ respectively. Following (8), with S_k and $S_{k+1/2}$ we fit two local linear regression models respectively denoted by $m_{k,1}(\widehat{x}^k + s, \widehat{\varepsilon}^{k,1})$ and $m_{k,2}(\widehat{x}^k + s, \widehat{\varepsilon}^{k,2})$. Then function values $f(\widehat{x}^k)$ and $f(\widehat{x}^k + s^k)$ are estimated by v_k and $v_{k+1/2}$ as follows.

$$v_k \;\triangleq\; c(\widehat{x}^k) + \mathbf{E}_{\widehat{\varepsilon}^{k,1}}[\, h(\widehat{x}^k, m_{k,1}(\widehat{x}^k, \widehat{\varepsilon}^{k,1})) \,+ g(\widehat{x}^k, m_{k,1}(\widehat{x}^k, \widehat{\varepsilon}^{k,1}))\,] \tag{13}$$

$$v_{k+1/2} \;\triangleq\; c(\widehat{x}^k + s^k) + \mathbf{E}_{\widehat{\varepsilon}^{k,2}}[\, h(\widehat{x}^k + s^k, m_{k,2}(\widehat{x}^k + s^k, \widehat{\varepsilon}^{k,2})) \,+ g(\widehat{x}^k + s^k, m_{k,2}(\widehat{x}^k + s^k, \widehat{\varepsilon}^{k,2}))] \tag{14}$$

The ratio of estimated actual-to-predicted reduction is then approximated as

$$\rho_k \triangleq (v_k - v_{k+1/2}) / \Big(f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k) \Big) \tag{15}$$

The criteria that whether the new step $\widehat{x}^k + s^k$ will be rejected or accepted contain two parts. One is that if the ratio ρ_k is large enough, the other is the that if the ratio between the generalized gradient $\chi_k(\widehat{x}^k)$ and the TR radius δ_k is large enough. In section 11.4 in [11], there is a discussion on how to compute the generalized gradient for a composite function $h(c(x))$ when h is a piecewise affine function and c is a differentiable function. We can easily extend the formulation of generalized

gradient for f_k under the assumption (A2). In particular, we have

$$\chi_k(x) = - \min_{\substack{x+d \in X \\ \|d\| \le 1}} f_k{}'(x;\, d)$$

$$= - \min_{\substack{x+d \in X \\ \|d\| \le 1}} \left(\begin{array}{l} \nabla c(x)^\top d + \dfrac{1}{N_k} \displaystyle\sum_{i=1}^{N_k} h'(x, m_k(x, e^{k,i});\, (d, (B_{1,k})^\top d)) \\ + \dfrac{1}{N_k} \displaystyle\sum_{i=1}^{N_k} \nabla_1 g(x, m_k(x, e^{k,i}))^\top d + \nabla_2 g(x, m_k(x, e^{k,i}))^\top (B_{1,k})^\top d \end{array} \right)$$

$$= - \min_{\substack{x+d \in X \\ \|d\| \le 1}} \left(\begin{array}{l} \nabla c(x)^\top d + \dfrac{1}{N_k} \displaystyle\sum_{i=1}^{N_k} \nabla_1 g(x, m_k(x, e^{k,i}))^\top d + \nabla_2 g(x, m_k(x, e^{k,i}))^\top (B_{1,k})^\top d \\ + \dfrac{1}{N_k} \displaystyle\sum_{i=1}^{N_k} \max_{j \in \mathcal{A}(x, e^{k,i})} \left\{ \nabla_1 h_j(x, m_k(x, e^{k,i}))^\top d + \nabla_2 h_j(x, m_k(x, e^{k,i})^\top (B_{1,k})^\top d \right\} \end{array} \right)$$

where $\mathcal{A}(x, e^{k,i}) \triangleq \{j \in \mathcal{J} : h_j(x, m_k(x, e^{k,i})) = h(x, m_k(x, e^{k,i}))\}$. Hence, the computation of generalized gradient is equivalent to solving a convex program as above.

2.4 CLEO algorithm

We present the CLEO algorithm which combines the learning step and the optimization step in Algorithm 1. At each iteration, given the data set T_k, the CLEO algorithm successively constructs an LLR model m_k and a trust region model f_k following lines 3-5 of Algorithm 1. A suitable step s^k is derived in line 6 satisfying the sufficient decrease condition (11). The actual-to-predicted reduction ratio ρ_k is computed through lines 7-9. According to the update rule from line 10 to line 14, a new iterate point is accepted if the sufficient function reduction is achieved, and the trust-region radius is small compared to the norm of the generalized gradient. The logic behind this update criterion on trust-region radius follows from the fact that the step size obtained by the minimization of a smooth function is typically proportional to the norm of its gradient, hence the trust region should be of comparable size as well. In the next iterate, we take the bandwidth of LLR model to be the same as the updated trust region radius. If the current local linear regression model does not fit the local dependency well, this would lead to a large ratio ρ_k, and then the suitable step is rejected so that the next LLR model with a shrunk bandwidth is more appropriate in a smaller region. In this way, the CLEO algorithm connects the predictive model with the optimization problem through a "smart" bandwidth choice. This connection automatically balances the trade-off between the exploration of the new decision region and the exploitation of the current decision region.

To practically implement CLEO, one needs to specify the size of data generated in each trust region. In the convergence analysis in section 3.2, we present the sample size requirement for the convergence. However, it is not a practical guide as most of the constants (i.e., Lipschitz gradients) in the sample size requirement are not usually available in practice. We assume that the sequence of sample sizes $\{N_k\}$ are given for now and we will discuss the empirical choices of sample sizes in Section 4.

Algorithm 1 CLEO

Initialization:
$\widehat{x}^0 \in X$, $\delta_0 \in (0, \delta_{\max})$ with $\delta_{\max} > 0$, $\gamma > 1$, $\eta_1 \in (0,1)$, $\eta_2 > 0$.
for $k = 0, 1, 2, \ldots$ **do**
  generate a set of samples $\{u^i\}$ from a uniform distribution $\mathcal{U}(\mathcal{B}(\mathbf{0}_p, 1))$. Let $\{x^i\} \triangleq \{\delta_k u^i + \widehat{x}^k\} \cap X$ and $\{\omega^i\}$ are corresponding scenarios according to the true models $\{m(x^i, \widetilde{\varepsilon})\}$. Let $T_k \triangleq \{(x^i, \omega^i)\}$.
  construct a local linear regression model $m_k(\widehat{x}^k + s, \widehat{\varepsilon}^k)$ by (8) with the data set T_k.
  construct the trust region model $f_k(\widehat{x}^k + s)$ by (9).
  compute a suitable step s^k of the TR subproblem (10) satisfying the condition (11).
  generate sets of data pairs S_k and $S_{k+1/2}$ uniformly in trust regions $\mathcal{B}(\widehat{x}^k, \|s_k\|/2) \cap X$ and $\mathcal{B}(\widehat{x}^k + s^k, \|s_k\|/2) \cap X$ respectively.
  construct two local linear regression model $m_{k,1}(\widehat{x}^k + s^k, \widehat{\varepsilon}^{k,1})$ and $m_{k,2}(\widehat{x}^k + s^k, \widehat{\varepsilon}^{k,2})$ by fitting data sets S_k and $S_{k+1/2}$ respectively.
  compute v_k by (13) , $v_{k+1/2}$ by (14), ρ_k by (15) and $\chi_k(\widehat{x}^k)$ by (12).
  if $\rho_k \geq \eta_1$ and $\chi_k(\widehat{x}^k) \geq \eta_2 \delta_k$ **then**
    $\widehat{x}^{k+1} = \widehat{x}^k + s^k$, $\delta_{k+1} = \min\ \{\gamma\delta_k,\ \delta\max\}$
  else
    $\widehat{x}^{k+1} = \widehat{x}^k$, $\delta_{k+1} = \gamma^{-1}\delta_k$
  end if
end for

3 Convergence analysis of the CLEO algorithm

In this section, we show that the sequence $\{\widehat{x}^k\}$ produced by the CLEO algorithm converges to a stationary point of (2) with probability 1. Intuitively, the convergence should hold if the trust region models $\{f_k\}$ are probabilistically accurate within the trust region. The condition of probabilistic accuracy that is required for the convergence is formally defined as probabilistically fully linearity in the literature [2, 9]. The convergence analysis for CLEO is comprised of two parts. In section 3.1, we first show that under the sample size requirement of data in trust regions, the TR random model f_k is probabilistically fully linear. Using this property, we then show in section 3.2 the sequential convergence of CLEO to a stationary point with probability 1. Though CLEO can be seen as an extension of the trust region method for unconstrained smooth optimization problem in [9], the convergence analysis for CLEO is nontrivial due to the challenges in dealing with a constrained, latently composite, nonsmooth and nonconvex stochastic optimization problem.

3.1 Probabilistically fully linear property of a TR model

In the following analysis, when we describe a random process in the CLEO algorithm, we use uppercase letters, such as the kth iterate $\widehat{X}^k$, to denote random variables, while we use lowercase letter to denote realizations of the random variable, such as $\widehat{x}^k$ which denotes the k-th iterate for a particular realization of our algorithm. This kind of notation will be applied to the LLR models $\{m_k\}, \{M_k\}$, TR random models $\{f_k\}, \{F_k\}$ and random value estimates $\{v_k, v_{k+1/2}\}, \{V_k, V_{k+1/2}\}$ as well. Hence, our algorithm results in a stochastic process $\{M_k, F_k, X_k, S_k, \Delta_k, V_k, V_{k+1/2}\}$. Our goal is to show that under certain conditions on the sequences $\{F_k\}$ and $\{V_k, V_{k+1/2}\}$, the resulting stochastic process has desirable convergence properties almost surely. We first provide concepts

of κ-fully linear approximation of deterministic and random trust region models respectively in Definition 2 and 3. The definition of fully linear model is introduced in section 6 of [12]. With this notion, model functions are characterized to behave similarly to Taylor approximations in a given trust region. This notion was generalized to random models in the probabilistic sense in [2].

Definition 2. Let $f : Z \to \mathbb{R}$ and $f_k : Z \to \mathbb{R}$ be locally Lipschitz continuous and directionally differentiable functions defined on an open set $Z \subseteq \mathbb{R}^m$ containing the convex compact set X. We say that the function f_k is a κ-fully linear model of f on $\mathcal{B}(\widehat{x}^k, \delta_k)$ with $\kappa = (\kappa_{ef}, \kappa_{ed})$ and $\kappa_{ef}, \kappa_{ed} \geq 0$, if for any $x \in \mathcal{B}(\widehat{x}^k, \delta_k) \cap X$,

$$\begin{aligned} \chi(x) - \chi_k(x) \;&\leq\; \kappa_{ed}\, \delta_k, \\ |\, f(x) - f_k(x)\,| &\leq \kappa_{ef}\, \delta_k^2, \end{aligned}$$

where $\chi(x)$ is defined in (5) and $\chi_k(x)$ is defined in (12).

Definition 3. Let $f : Z \to \mathbb{R}$ be a locally Lipschitz continuous and directional differentiable function defined on an open set $Z \subseteq \mathbb{R}^m$ containing the convex compact set X. Let $\{F_k : Z \to \mathbb{R}\}$ be a sequence of random functions, each of which is locally Lipschitz continuous and directionally differentiable almost surely. A sequence of random functions $\{F_k\}$ are said to be α-probabilistically κ-fully linear with respect to $\{\mathcal{B}(\widehat{X}^k, \Delta_k)\}$ if and only if for a scalar $\alpha \in (0, 1)$, a constant vector $\kappa \in \mathbb{R}^2_+$ and any $k \geq 1$, the event $I_k = \{\, F_k$ is a κ-fully linear model of f on $\mathcal{B}(\widehat{X}_k, \Delta_k)\,\}$ satisfies the condition

$$\mathbb{P}\,(\, I_k \,|\, \mathcal{F}_{k-1}\,) \;\geq\; \alpha,$$

where $\mathcal{F}_{k-1}$ denotes the σ-algebra generated by $\{F_i, V_i, V_{i+1/2}\}_{i=0}^{k-1}$, and $\mathbb{P}(\bullet \,|\, \mathcal{F}_{k-1})$ is a conditional probability given the past history of $\mathcal{F}_{k-1}$.

In addition to the probabilistically accurate condition for TR random models, we require sufficiently accurate value estimates.

Definition 4. The value estimates v_k and $v_{k+1/2}$ are ϵ_F-accurate estimates of $f(\widehat{x}^k)$ and $f(\widehat{x}^k + s^k)$, respectively, for a given δ_k if and only if

$$|\, v_k - f(\widehat{x}^k)\,| \;\leq\; \epsilon_F\, \delta_k^2, \quad |\, v_{k+1/2} - f(\widehat{x}^k + s^k)\,| \;\leq\; \epsilon_F\, \delta_k^2.$$

Definition 5. A sequence of random value estimates $\{(V_k, V_{k+1/2})\}$ are said to be β-probabilistically ϵ_F-accurate with respect to the corresponding sequence $\{X_k, \Delta_k, S_k\}$ if for the nonnegative constants $\beta \in [0, 1]$, ϵ_F and any $k \geq 1$, the event

$$J_k = \{\, V_k, V_{k+1/2} \text{ are } \epsilon_F\text{-accurate estimates of } f(\widehat{x}^k) \text{ and } f(\widehat{x}^k + s^k) \text{ respectively for } \Delta_k\,\}$$

satisfies the condition

$$\mathbb{P}\,(\, J_k \,|\, \mathcal{F}_{k-1/2}\,) \geq \beta,$$

where $\mathcal{F}_{k-1/2}$ denotes the σ-algebra generated by $\{F_i\}_{i=1}^{k}$ and $\{V_i, V_{i+1/2}\}_{i=0}^{k-1}$, and $\mathbb{P}(\bullet \,|\, \mathcal{F}_{k-1/2})$ is a conditional probability given the past history of $\mathcal{F}_{k-1/2}$.

For the linear regression model, we introduce definitions of poised and strongly Λ-poised condition according to section 6 of [12], which will be utilized for the probabilistically fully linear property

and probabilistically accurate property. Recall that a trust region data set of size N_k is denoted by $T_k = \{(x^i, \omega^i)\}$. Let

$$\widehat{U}^k \triangleq \begin{pmatrix} 1 & (x^1)^\top \\ \vdots & \vdots \\ 1 & (x^{N_k})^\top \end{pmatrix}$$

Definition 6. The set $\{x^i\}$ is poised for the linear regression model if the matrix $\widehat{U}^k$ has full column rank.

If the set $\{x^i\}$ is poised, then the least-square linear regression is unique. However, the condition of poisedness is not sufficient to derive a uniform error bounds for the convergence analysis. Under the condition of poisedness, when the number of samples is allowed to grow arbitrarily large, we introduce the strongly Λ-poisedness condition which imposes an upper bound on the Lagrangian polynomials $l(z)$ in the ℓ_2 norm.

Definition 7. Let a scalar $\Lambda > 0$ and a set $\mathcal{B} \subseteq \mathbb{R}^p$ be given. A poised set $\{x^i\}$ of size N_k is said to be strongly Λ-poised for the linear regression model in $\mathcal{B}$ if $\sqrt{N_k}\ \max_{z\in\mathcal{B}} \| l(z) \| \leq \Lambda$, where the Lagrangian polynomials $l(z) \triangleq (\widehat{U}^k)((\widehat{U}^k)^\top \widehat{U}^k)^{-1} \begin{pmatrix} 1 \\ z \end{pmatrix}$.

We refer to section 6.5 in [12] on ensuring a strongly Λ-poised set for derivative-free optimization. Specifically, for the regression model, algorithm 6.7 in [12] provides an approach to compute a Λ-poised set given an initial set of points Y based on Theorem 6.3 and Algorithm 6.3 in [12]. To show that trust region models $\{F_k\}$ constructed in (10) satisfies the α-probabilistically κ-fully linear property for $\alpha \in (0, 1)$, we need the probabilistically accurate properties of the LLR in Proposition 8 with the proof given in the Appendix. In [26], similar properties are analyzed for regression models under the strongly Λ-poised condition using the Markov inequality. Our analysis of this property is different from theirs by using the Law of Large Numbers and Berry-Esseen Theorem under the same strongly Λ-poised condition.

Proposition 8. Under assumptions (A1) and (A3), given a feasible point $\widehat{x}_k \in X$ and a trust region radius $\delta_k > 0$, for the data set $T_k = \{(x^i, \omega^i)\}$ of size N_k generated from line 3 in CLEO, suppose that the set $\{x^i\}$ is strongly Λ-poised in $\mathcal{B}(\widehat{x}^k, \delta_k) \cap X$. Let $\widehat{B}^{k,1}$ and $\widehat{B}^{k,0}$ be the least square estimators of the local linear regression model following (6). Then for any $z \in \mathcal{B}(\widehat{x}^k, \delta_k) \cap X$, we have the following results.

(a) For any $0 < \alpha < 1$, there exists a constant $\kappa_{eg} > 0$ such that when $N_k \geq \max\{O(\delta_k^{-4}\, \kappa_{eg}^{-2}),\, O(\alpha^{-2})\}$,

$$\mathbb{P}\left(\| \nabla\psi(x) - \widehat{B}^{k,1} \| \,\leq\, \kappa_{eg}\, \Delta_k,\, \forall x \in \mathcal{B}(\widehat{X}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1} \right) \geq 1 - \alpha.$$

(b) For any $0 < \alpha < 1$, there exists a constant $\kappa_{ef} > 0$, when $N_k \geq O(\delta_k^{-4}\kappa_{ef}^{-2}\alpha^{-1})$,

$$\mathbb{P}\left(\| \psi(x) - (\widehat{B}^{k,1})^\top x - (\widehat{B}^{k,0})^\top \| \,\leq\, \kappa_{ef}\, \Delta_k^2,\, \forall x \in \mathcal{B}(\widehat{X}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1} \right) \geq 1 - \alpha.$$

Now we establish the probabilistically fully linear property in Proposition 9 with the proof given in the Appendix. It relates the probabilistic error of TR models with the probabilistic error of

LLR estimators given in Proposition 8. It is worth noting that we utilize some properties of ϵ-approximate directional derivative in the proof to provide the probabilistic bound for the gap $\chi(x) - \chi_k(x)$.

Proposition 9. Suppose assumptions (A1)-(A4) hold for the composite SP problem (2). At the k-th iteration of the CLEO algorithm, given a feasible point $\widehat{x}_k \in X$ and a trust region radius $\delta_k > 0$, for the data set $T_k = \{(x^i, \omega^i)\}$ of size N_k generated in CLEO, suppose that the set $\{x^i\}$ is strongly Λ-poised in $\mathcal{B}(\widehat{x}^k, \delta_k) \cap X$.

(a) There exists a constant $\kappa_{ef} > 0$, when $N_k \geq \max\{O(\delta_k^{-4}\kappa_{ef}^{-4}),\, O(\alpha^{-2})\}$, we have

$$\mathbb{P}\left(|f(x) - F_k(x)| \geq \kappa_{ef}\Delta_k^2, \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1}\right) \leq \frac{\alpha}{2}.$$

(b) There exists a constant $\kappa_{ed} > 0$, when $N_k \geq \max\{O(\delta_k^{-2}\kappa_{ed}^{-2}),\, O(\alpha^{-2})\}$, we have

$$\mathbb{P}\left(\chi(x) - \chi_k(x) \geq \kappa_{ed}\Delta_k, \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1}\right) \leq \frac{\alpha}{2}.$$

From (a) and (b), we can derive that there exists a constant vector $\kappa = (\kappa_{ef}, \kappa_{ed})$, when $N_k \geq O(\max\{\delta_k^{-4}, O(\alpha^{-2})\})$, we have

$$\mathbb{P}\left(|\, f(x) - F_k(x)\,| \,\leq\, \kappa_{ef}\,\Delta_k^2,\ \chi(x) - \chi_k(x) \,\leq\, \kappa_{ed}\,\Delta_k,\ \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \,\middle|\, \mathcal{F}_{k-1}\right) \geq 1 - \alpha.$$

We can derive the probability accuracy of the random estimates $(V_k, V_{k+1/2})$ by using the similar analysis as in Proposition 9. So we omit the proof and present the result as follows.

Proposition 10. *For the composite SP problem* (2) *under assumptions (A1)- (A4), for the data set $T_k = \{(x^i, \omega^i)\}$ of size N_k generated in CLEO, suppose that the set $\{x^i\}$ is strongly Λ-poised in $\mathcal{B}(\widehat{x}^k, \delta_k) \cap X$. Then for any $0 < \beta < 1$, there exist positive constants ϵ_F, and $M_k(\beta)$ of order $O(\Delta_k^{-4})$, such that if $|S_k|, |S_{k+1/2}| \geq M_k(\beta)$, then random estimates $(V_k, V_{k+1/2})$ is β-probabilistically ϵ_F-fully accurate with respect to $B(\widehat{X}^k, \Delta_k)$.*

3.2 Convergence analysis to a d-stationary point

The convergence analysis serves as an extension of trust region method with random models [9] to the nonsmooth, composite, constrained stochastic programming problem. The convergence analysis is based on Proposition 9 and Proposition 10, which hold under the requirement of sample size in the trust region with the scale of constants depending on the smooth functions $\{h_j(\cdot)\}_{j\in\mathcal{J}}$ and $g(\cdot)$. With a fair estimation of the constants, we assume that in the process of the CLEO algorithm, we could generate data pairs from reality or simulator so that the size of TR data sets $\{T_k\}$ satisfy the requirements in Proposition 9 and Proposition 10. While this requirement may be demanding sometimes, numerical experiments in section 4 show that CLEO needs much smaller size of data than the PO scheme in achieving a suboptimal objective value. Hence, we assume the following assumption on the sample size requirement for the sake of the theoretical convergence analysis.

(**B**) Suppose for some chosen probability parameters α and β, for any given positive number k, feasible point $\widehat{x}^k$ and trust region radius δ_k, we are able to generate a sample set of size in order of $O(\delta_k^{-4})$ in the trust region $B(\widehat{x}^k, \delta_k) \cap X$ with which the TR random model F_k is α-probabilistically κ-fully linear and random value estimates $(V_k, V_{k+1/2})$ are β-probabilistically ϵ_F-accurate.

The following lemmas hold under the assumptions (A1)-(A4). So for the sake of brevity, we will not include the assumptions in the lemmas. We first show the guarantees of the decrease of the objective value in the composite SP problem (2).

Lemma 11. Suppose that the TR model f_k is $\kappa = (\kappa_{ef}, \kappa_{ed})$-fully linear on $\mathcal{B}(\widehat{x}^k, \delta_k)$ with $\delta_k \leq \delta_{\max}$. If

$$\delta_k \leq \frac{1}{\left(\frac{4\kappa_{ef}}{\kappa_{dcp}} + \kappa_{ed}\right)\delta_{\max}} \chi(\widehat{x}^k), \tag{16}$$

then a suitable step s_k under (11) leads to an improvement in $f(\widehat{x}^k + s^k)$ such that

$$f(\widehat{x}^k + s^k) - f(\widehat{x}^k) \leq -C_1 \chi(\widehat{x}^k)\delta_k.$$

where $C_1 \triangleq \dfrac{2\kappa_{dcp}\,\kappa_{ef}}{(4\kappa_{ef} + \kappa_{ed}\,\kappa_{dcp})\,\delta_{\max}}$.

Proof. The definition of a $\kappa = (\kappa_{ef}, \kappa_{ed})$-fully linear model yields that,

$$\chi_k(\widehat{x}^k) \geq \chi(\widehat{x}^k) - \kappa_{ed}\delta_k \geq \frac{4\kappa_{ef}}{\kappa_{dcp}}\,\delta_k \tag{17}$$

$$\chi_k(\widehat{x}^k) \geq \chi(\widehat{x}^k) - \kappa_{ed}\delta_k \geq \left(1 - \frac{\kappa_{ed}\,\kappa_{dcp}}{4\kappa_{ef} + \kappa_{ed}\,\kappa_{dcp}}\right)\chi(\widehat{x}^k) = \frac{4\kappa_{ef}}{4\kappa_{ef} + \kappa_{ed}\,\kappa_{dcp}}\chi(\widehat{x}^k) \tag{18}$$

With a suitable step s^k by (11)

$$f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k) \;\geq\; \kappa_{dcp}\,\chi_k(\widehat{x}^k)\,\min\{\delta_k, 1\} \geq \frac{\kappa_{dcp}}{\delta_{\max}}\,\chi_k(\widehat{x}^k)\,\delta_k,$$

From the definition of a κ-fully linear model, we have

$$\begin{aligned} f(\widehat{x}^k + s^k) - f(\widehat{x}^k) &= f(\widehat{x}^k + s^k) - f_k(\widehat{x}^k + s^k) + f_k(\widehat{x}^k + s^k) - f_k(\widehat{x}^k) + f_k(\widehat{x}^k) - f(\widehat{x}^k) \\ &\leq 2\,\kappa_{ef}\,\delta_k^2 - \frac{\kappa_{dcp}}{\delta_{\max}}\,\chi_k(\widehat{x}^k)\,\delta_k \\ &\leq -\frac{\kappa_{dcp}}{2\,\delta_{\max}}\,\chi_k(\widehat{x}^k)\,\delta_k, \end{aligned} \tag{19}$$

where the last inequality (19) is derived from (17) that $\delta_k \leq \frac{\kappa_{dcp}}{4\,\kappa_{ef}\cdot\delta_{\max}}\chi_k(\widehat{x}^k)$. Because of (18), we have

$$f(\widehat{x}^k + s^k) - f(\widehat{x}^k) \leq -\frac{2\kappa_{dcp}\,\kappa_{ef}}{(4\kappa_{ef} + \kappa_{ed}\,\kappa_{dcp})\,\delta_{\max}}\,\chi(\widehat{x}^k)\delta_k$$

□

Lemma 12. Suppose that the TR model f_k is κ-fully linear on $B(\widehat{x}^k, \delta_k)$ and the estimates $(v_k, v_{k+1/2})$ is ϵ_F-accurate with $\kappa = (\kappa_{ef}, \kappa_{ed})$ and $\epsilon_F \leq \kappa_{ef}$. If $\delta_k \leq \delta_0$, and

$$\delta_k \;\leq\; \min\left\{\frac{1}{\eta_2},\, \frac{\kappa_{dcp}(1-\eta_1)}{4\kappa_{ef}\,\delta_{\max}}\right\}\chi_k(\widehat{x}^k).$$

where the constant η_2 is the step acceptance parameter in line 10 of the CLEO algorithm, then k-th iteration is successful, i.e., the sufficient decrease condition $\rho_k \geq \eta_1$ defined in (15) is achieved.

Proof. The trust region random function f_k being $(\kappa_{ef}, \kappa_{eg})$-fully linear implies that

$$| f(\widehat{x}^k) - f_k(\widehat{x}^k) | \le \kappa_{ef}\, \delta_k^2, \quad \text{and} \quad | f(\widehat{x}^k + s^k) - f_k(\widehat{x}^k + s^k) | \le \kappa_{ef}\, \delta_k^2. \tag{20}$$

Since value estimates $(v_k, v_{k+1/2})$ are ϵ_F-accurate with $\epsilon_F \le \kappa_{ef}$, we have

$$| v_k - f(\widehat{x}^k) | \le \kappa_{ef} \delta_k^2, \quad \text{and} \quad | v_{k+1/2} - f(\widehat{x}^k + s^k) | \le \kappa_{ef}\, \delta_k^2. \tag{21}$$

Consider the parameter

$$\begin{aligned}
\rho_k &= \frac{v_k - v_{k+1/2}}{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)} \\
&= \frac{v_k - f(\widehat{x}^k)}{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)} + \frac{f(\widehat{x}^k) - f_k(\widehat{x}^k)}{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)} + \frac{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)}{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)} \\
&\quad + \frac{f_k(\widehat{x}^k + s^k) - f(\widehat{x}^k + s^k)}{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)} + \frac{f(\widehat{x}^k + s^k) - v_{k+1/2}}{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)}.
\end{aligned}$$

Then

$$\begin{aligned}
| \rho_k - 1 | \le\ & \frac{|v_k - f(\widehat{x}^k)|}{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)} + \frac{|f(\widehat{x}^k) - f_k(\widehat{x}^k)|}{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)} \\
& + \frac{|f_k(\widehat{x}^k + s^k) - f(\widehat{x}^k + s^k)|}{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)} + \frac{|f(\widehat{x}^k + s^k) - v_{k+1/2}|}{f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k)}
\end{aligned}$$

By the condition of a suitable step in (11), and inequalities (20) and (21), we have

$$| \rho_k - 1 | \le \frac{4\, \kappa_{ef}\, \delta_{\max}\, \delta_k}{\kappa_{dcp} \chi_k(\widehat{x}^k)} \le 1 - \eta_1,$$

where the last inequality comes from the fact that $\delta_k \le \frac{\kappa_{dcp}(1-\eta_1)}{4\kappa_{ef}} \delta_{\max} \chi_k(\widehat{x}^k)$. □

Lemma 13. Suppose the function value estimates $\{(v_k, v_{k+1/2})\}$ are ϵ_F-accurate and

$$\epsilon_F < \frac{1}{2}\, \eta_1\, \eta_2\, \frac{\kappa_{dcp}}{\delta_{\max}}.$$

If the kth iteration is successful, i.e. a trial step s^k is accepted, then the improvement in f is bounded such that

$$f(\widehat{x}^{k+1}) - f(\widehat{x}^k) \le -C_2\, \delta_k^2,$$

where $C_2 \triangleq \eta_1\, \eta_2\, \frac{\kappa_{dcp}}{\delta_{\max}} - 2\epsilon_F$.

Proof. When k-th iteration is successful, $\rho_k \ge \eta_1$ and $\chi_k(\widehat{x}^k) \ge \eta_2\, \delta_k$, then

$$\begin{aligned}
& v_k - v_{k+1/2} \\
& \ge\ \eta_1 \left(f_k(\widehat{x}^k) - f_k(\widehat{x}^k + s^k) \right) \ge \eta_1\, \kappa_{dcp}\, \chi_k(\widehat{x}^k) \min\{\delta_k, 1\} \ge \eta_1\, \eta_2\, \frac{\kappa_{dcp}}{\delta_{\max}}\, \delta_k^2.
\end{aligned}$$

Since v_k and $v_{k+1/2}$ are ϵ_F-accurate, the improvement of f can be bounded as below.

$$f(\widehat{x}^k) - f(\widehat{x}^k + s^k) = \Big(f(\widehat{x}^k) - v_k \Big) + \big(v_k - v_{k+1/2} \big) + \Big(v_{k+1/2} - f(\widehat{x}^k + s^k) \Big)$$
$$\geq \left(\eta_1 \, \eta_2 \, \frac{\kappa_{dcp}}{\delta_{\max}} - 2\epsilon_F \right) \delta_k^2.$$

□

Now with Lemma 11, 12, 13, we show in Proposition 14 that the probability parameters α and β can be chosen as constant values, so that the sum of trust region radii in the CLEO algorithm is finite almost surely. The specific condition for parameters α and β is illustrated in the proof of Proposition 14 in the Appendix. Furthermore, we show that the sequence of random iterates converges to a directional stationary point almost surely in Theorem 15 with the proof in the Appendix. It is worth noting that due to the constrained, composite nonsmooth trust region models, some modifications of the proofs are made based on proof of Theorem 4.11 in [9].

Proposition 14. For the composite SP (2), under the assumptions (A1)-(A4), suppose that the strongly Λ-poisedness condition holds for data sets in every trust region, and the step acceptance parameter η_2 and the accuracy parameter ϵ_F satisfy

$$\eta_2 \geq \frac{4\kappa_{ef}}{\kappa_{dcp}(1-\eta_1)}, \text{ and } \epsilon_F \leq \min\left\{ \kappa_{ef}, \frac{1}{4}\,\eta_1\,\eta_2\,\frac{\kappa_{dcp}}{\delta_{\max}} \right\}. \tag{22}$$

Then the probability parameters α and β can be chosen so that if the assumption (B) holds for these values, then the sequence of trust-region radius, $\{\Delta_k\}$ generated by the CLEO algorithm satisfies $\sum_{k=0}^{\infty} \Delta_k^2 < \infty$ almost surely.

Theorem 15. For the composite SP (2), under the same assumptions as in Proposition 14, the probability parameters α and β can be chosen so that if the assumption (B) holds for these values, we have $\lim_{k\to\infty} \chi(\widehat{X}^k) = 0$ almost surely; moreover, the sequence of random iterates $\{X^k\}$ generated by the CLEO Algorithm, converges to a directional stationary point almost surely.

4 Numerical experiments

In this section, we discuss the performance of the CLEO algorithm compared with the PO scheme. In the PO scheme, we first predict the dependency of the uncertainty on decision variables using statistical learning models and then solve the SP problem equipped with prediction models. In order to make a fair comparison of the PO scheme and the CLEO algorithm, we use the L-BFGS method to solve both the approximate optimization problems in the PO scheme or the trust region subproblems in the CLEO scheme. The comparisons are conducted for three decision-making problems, including a nonconvex problem on synthetic data, a joint production and pricing problem on synthetic data and a price sensitive problem of hotel rooms on real world data. The code of the CLEO algorithm for the joint production and pricing instance is available at *github.com/junyiliu99/CLEO*.

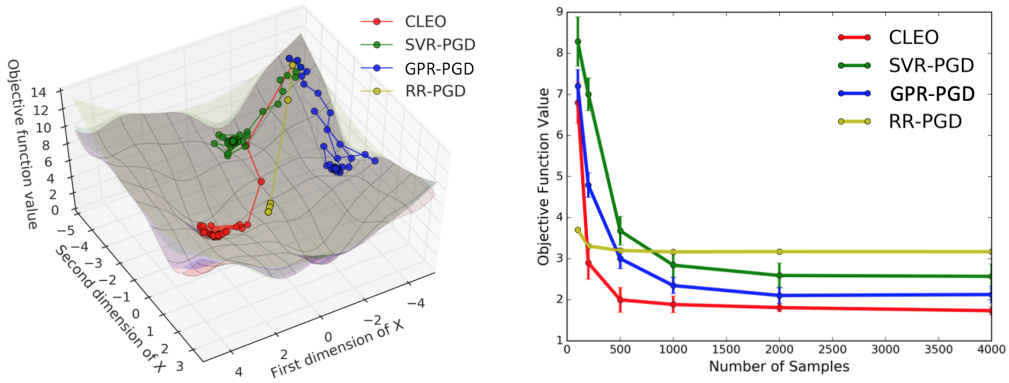
4.1 A synthetic nonconvex problem

We consider an SP problem as follows.

$$\underset{\substack{-4 \leq x_1 \leq 4 \\ -5 \leq x_2 \leq 3}}{\text{minimize}} \quad \frac{1}{5} \, \| \mathbf{x} \|^2 + \mathbf{E}_{\widetilde{\omega}|\mathbf{x}}[\, \widetilde{\omega}^2 \,] \tag{23}$$

The random variable $\widetilde{\omega}$ is a regression model of the decision variable $\mathbf{x}$, i.e., $\widetilde{\omega} = -\sin(x_1)+\sin(x_2)+\widetilde{\varepsilon}$, and $\widetilde{\varepsilon} \sim \mathcal{N}(0,1)$. The SP problem (23) is thus a nonconvex and smooth problem. Without the knowledge of the actual relationship of $\widetilde{\omega}$ and $\mathbf{x}$, we utilize the CLEO algorithm as well as the PO scheme with prediction models including Ridge Regression(RR), Support Vector Regression(SVR), and Gaussian Process(GP) with randomly generated data pairs to seek a near-optimal decision of problem (23) by proximal gradient descent (PGD). The hyperparameter tuning of these prediction models is accompalished by grid search, which means that we manually discretize the value space of hyperparameters and select the best hyperparameters in the prediction model. The parameters of prediction models are estimated by the package "scikit_learn" in Python. When implementing the CLEO algorithm, the data in each neighborhood are actively generated and the data size is set to be a constant number (e.g., 10). The performance of four algorithms is evaluated with respect to two criteria: 1) a snapshot of convergence process in Figure 1(a) where red, green, blue and yellow surfaces respectively refer to the surface of ground truth objective function, objective value with the uncertainty fitted by SVR, GPR and RR; 2) objective value versus the number of samples in Figure 1(b) with the same stopping criterion. These plots are produced based on a fair and careful parameter-tuning together with an average of 50 random replications.

As expected, CLEO converges to a stationary point of the true objective surface with the smallest number of iterations. As indicated by Figure 1(a), the yellow surface constructed by RR model does not fit well to the true one, therefore, PO scheme with RR model does not provide the convergence to any stationary point. When the sample size is large, the surface constructed by GPR is close to the true surface. As shown in Figure 1(b), the solution produced by the PO scheme with GPR has its objective value slightly higher than the one produced by CLEO. However, the derived objective values over replications in the PO scheme with GPR are more volatile than the CLEO scheme. Moreover, training a GPR model is very time-consuming when the sample size grows large.



(a) Process snapshot

(b) Objective value

Figure 1: Comparisons of the CLEO algorithm and the PO schemes for a synthetic nonconvex problem

4.2 Joint production and pricing problem

We follow the setting in the joint pricing and production example in Section 1.1 with $K = 1$, demand curve $\phi(p) = a\frac{e^{b-cp}}{1+e^{b-cp}}$ and the noise $\widetilde{\varepsilon} \sim U(-\bar{\varepsilon}, \bar{\varepsilon})$ and parameters $a = 470, b = 6, c =$

$0.3, \bar{\varepsilon} = 0.1$. We evaluate the performance of the CLEO algorithm by choosing different sample sizes in trust regions in Figure 2. As we can see, except for the sample size of 2, CLEO algorithms with other choices of sample size have similar performance in terms of the convergence of the objective value. This indicates that the sample size requirement in the assumption (B) is for the sake of the theoretical analysis, whereas in practice the sample size in the neighborhoods could be set as a constant in the CLEO algorithm.

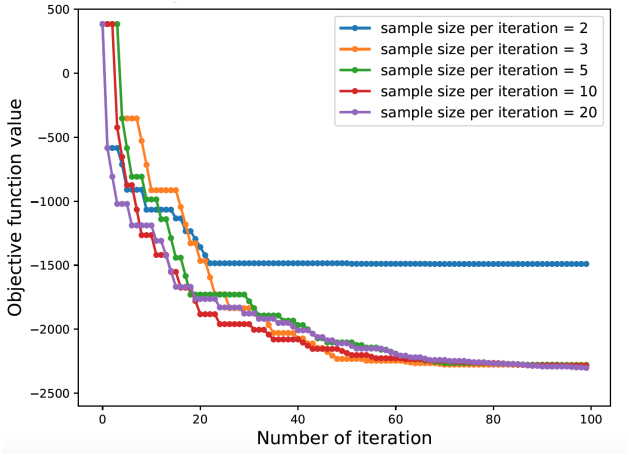


Figure 2: Convergence curve of CLEO for the sample size chosen as 2,3,5,10,20

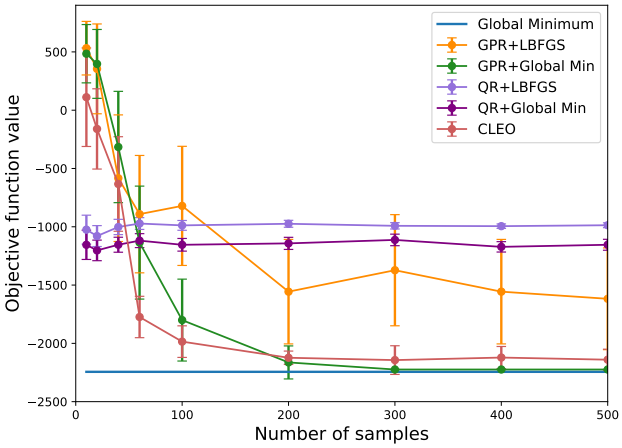


Figure 3: Convergence curve of the CLEO algorithm and the PO schemes for the joint production and pricing problem.

We compare the performance of the CLEO algorithm with the PO scheme with two prediction models, GPR model and Quadratic Regression (QR) model. The blue line at the bottom of Figure 3 represents the globally minimal objective value of the true composite SP problem (3), which

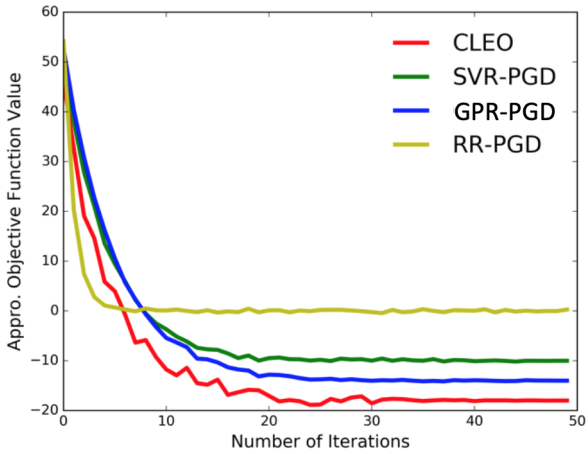


Figure 4: Convergence curve for the price-senstive problem of hotel rooms

could be taken as the ground truth. The red curve represents the objective function value at the solution derived by the CLEO algorithm with respect to the number of samples. The red curve indicates that the CLEO algorithm converges to the value slightly higher than the ground truth as the sample size increases.

The performance of the PO schemes relies on the accuracy of the prediction models and the optimality of the solution derived by the solver. To illustrate how these two parts contribute to the overall performance of the PO schemes, we compute two types of convergence curves for each PO scheme. For the approximate optimization problem composed with the GPR model, the green curve represents the globally optimal objective value and the orange curve represents the objective value at the solution derived by the L-BFGS solver. The green curve indicates that its global minimum converges to the ground truth as the sample size increases, whereas the derived solution by L-BFGS solver for the PO scheme is suboptimal with a large variance indicated by the orange curve. Since we only consider two decision variables in this instance, the global minimum of the highly nonconvex composite problem can be found graphically. However, when the dimension is higher than 2, the gap between the orange curve and the ground truth is mainly the result of the highly nonconvex structure of the approximate optimization problem composed with the GPR model. For the approximate optimization problem composed with the QR model, the gap between the light and dark purple curves is small and relatively stable, whereas the gap between the purple curves and the ground truth is large. This indicates that the estimation error of the latent dependency using QR models results in an unsatisfactory solution even though the composite optimization problem in the PO scheme can be solved efficiently. Combining the performance of two PO schemes, we can see the trade-off between the accuracy of the prediction model and the complexity of the composite optimization problem.

4.3 A pricing problem for hotel rooms

We consider a pricing problem for hotel rooms. We use the hotel booking dataset from [7] which contains 17,838 booking records with check-in dates from March 12, 2007 to April 15, 2007, in one of five continental U.S. hotels. The decision variable $\mathbf{x} \in \mathbb{R}^3_+$ is the price vector corresponding to

three room types: King, Queen, Standard, and the demand vector $\widetilde{\omega}$ is the number of booking records for each room type. The decision-making problem is formulated as

$$\underset{\mathbf{x}\in\mathbb{R}^3_+}{\text{maximize}} \quad \mathbf{E}_{\widetilde{\omega}|\mathbf{x}}\left[\,\mathbf{x}^T\widetilde{\omega}\,\right] - \lambda \parallel \mathbf{x} \parallel^2 \tag{24}$$

where $\lambda \geq 0$ is a parameter for the regularization of the price. Regardless of other market factors, the demand ω is implicitly affected by the price decision $\mathbf{x}$. We use the CLEO algorithm and three PO schemes with SVR, GP and RR for computing the price decision. Since the true optimization problem is unknown, the objective value at a solution is estimated by the sample average using samples within a neighborhood to the solution. In Figure 4, because of the randomness of the TR model constructed in CLEO, the performance curve mildly fluctuates, but overall it dominates the other three approaches in terms of the rate of progress and the objective value for this data set.

5 Conclusions

We proposed the CLEO algorithm coupling the local regression models and trust region methods for solving the stochastic programming problem with the decision-dependent uncertainty. The CLEO algorithm iteratively deals with local behaviors in both estimation and optimization problems to control the overall complexity with the asymptotic convergence to a stationary point with probability 1. The CLEO algorithm is then supported by computational results in simulation experiments and a real-world pricing problem. It shows that the CLEO algorithm outperforms the the uncoupled "Predict then Optimize" approach in terms of the convergent objective value and the associated variance.

Acknowledgments.

The first and the third authors acknowledge the support of AFOSR under Grant FA9550-15-1-0267 and FA9550-20-1-0006.

References

[1] Victor F Araman and René Caldentey. Dynamic pricing for nonperishable products with demand learning. *Operations research*, 57(5):1169–1188, 2009.

[2] Afonso S Bandeira, Katya Scheinberg, and Luís N Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM Journal on Optimization*, 24(3):1238–1264, 2014.

[3] Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 2019.

[4] Omar Besbes and Assaf Zeevi. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

[5] John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[6] Jose Blanchet, Coralia Cartis, Matt Menickelly, and Katya Scheinberg. Convergence rate analysis of a stochastic trust region method for nonconvex optimization. *arXiv preprint arXiv:1609.07428*, 2016.

[7] Tudor Bodea, Mark Ferguson, and Laurie Garrow. Data set–choice-based revenue management: Data from a major hotel chain. *Manufacturing & Service Operations Management*, 11(2):356–361, 2009.

[8] Boxiao Chen, Xiuli Chao, and Hyun-Soo Ahn. Coordinating pricing and inventory replenishment with nonparametric demand learning. *Operations Research*, 67(4):1035–1052, 2019.

[9] Ruobing Chen, Matt Menickelly, and Katya Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169(2):447–487, 2018.

[10] Wang Chi Cheung, David Simchi-Levi, and He Wang. Dynamic pricing and demand learning with limited price experimentation. *Operations Research*, 65(6):1722–1731, 2017.

[11] Andrew R Conn, Nicholas IM Gould, and Ph L Toint. *Trust region methods*, volume 1. SIAM, 2000.

[12] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*, volume 8. SIAM, 2009.

[13] William L Cooper, Tito Homem-de Mello, and Anton J Kleywegt. Models of the spiral-down effect in revenue management. *Operations Research*, 54(5):968–987, 2006.

[14] Arnoud V den Boer. Dynamic pricing and learning: historical origins, current research, and new directions. *Surveys in operations research and management science*, 20(1):1–18, 2015.

[15] John C Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

[16] Jitka Dupačová. Optimization under exogenous and endogenous uncertainty. *University of West Bohemia in Pilsen*, 2006.

[17] Rick Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

[18] Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *arXiv preprint arXiv:1710.08005*, 2017.

[19] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[20] Vikas Goel and Ignacio E Grossmann. A class of stochastic programs with decision dependent uncertainty. *Mathematical programming*, 108(2-3):355–394, 2006.

[21] Serge Gratton, Clément W Royer, Luís N Vicente, and Zaikun Zhang. Complexity and global rates of trust-region methods based on probabilistic models. *IMA Journal of Numerical Analysis*, 38(3):1579–1597, 2018.

[22] Holger Heitsch and Werner Römisch. Scenario reduction algorithms in stochastic programming. *Computational optimization and applications*, 24(2-3):187–206, 2003.

[23] Lars Hellemo, Paul I Barton, and Asgeir Tomasgard. Decision-dependent probabilities in stochastic programs with recourse. *Computational Management Science*, 15(3-4):369–395, 2018.

[24] Julia L Higle and Suvrajeet Sen. Finite master programs in regularized stochastic decomposition. *Mathematical Programming*, 67(1-3):143–168, 1994.

[25] Michal Kaut and Stein W Wallace. Evaluation of scenario-generation methods for stochastic programming. *Pacific Journal of Optimization*, 3(2):257–271, 2007.

[26] Jeffrey Larson and Stephen C Billups. Stochastic derivative-free optimization using a trust region framework. *Computational Optimization and applications*, 64(3):619–645, 2016.

[27] Jeffrey Larson, Matt Menickelly, and Stefan M Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019.

[28] Soonhui Lee, Tito Homem-de Mello, and Anton J Kleywegt. Newsvendor-type models with decision-dependent uncertainty. *Mathematical Methods of Operations Research*, 76(2):189–221, 2012.

[29] Jiajin Li, Anthony Man-Cho So, and Wing-Kin Ma. Understanding notions of stationarity in non-smooth optimization. *arXiv preprint arXiv:2006.14901*, 2020.

[30] Junyi Liu and Suvrajeet Sen. Asymptotic results of stochastic decomposition for two-stage stochastic quadratic programming. *SIAM Journal on Optimization*, 30(1):823–852, 2020.

[31] Liwan H. Liyanage and J. George Shanthikumar. A practical inventory control policy using operational statistics. *Operations Research Letters*, 33(4), 2005.

[32] Clive Loader. *Local regression and likelihood*. Springer Science & Business Media, 2006.

[33] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[34] Nilay Noyan, Gábor Rudolf, and Miguel Lejeune. Distributionally robust optimization with decision-dependent ambiguity set. *Optimization Online Http://www.optimization-online.org/DBHTML/2018/09/6821. html*, 2018.

[35] Srinivas Peeta, F Sibel Salman, Dilek Gunnec, and Kannan Viswanath. Pre-disaster investment decisions for strengthening a highway network. *Computers & Operations Research*, 37(10):1708–1719, 2010.

[36] Sashank J Reddi, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems*, pages 1145–1153, 2016.

[37] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[38] Suvrajeet Sen. Stochastic programming. In *Encyclopedia of Operations Research and Management Science*, pages 1486–1497. Springer, 2013.

[39] Suvrajeet Sen and Yunxiao Deng. Learning enabled optimization: Towards a fusion of statistical learning and stochastic optimization, 2017.

[40] Suvrajeet Sen and Yifan Liu. Mitigating uncertainty via compromise decisions in two-stage stochastic linear programming: Variance reduction. *Operations Research*, 64(6):1422–1437, 2016.

[41] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

6 Appendix

Lemma 16 (Berry Esseen Theorem)**.** Let $X_1, X_2, ...,$ be identitcal independent $\mathbb{R}^d$-valued random vectors with $\mathbf{E}[X_1] = 0$ and $\mathrm{VaR}(X_1) = \Sigma$ assuming that Σ is invertible. Let $Z \sim N(0, \Sigma)$ be a d-dimensional Gaussian random vector. Then for all convex sets $S \in \mathbb{R}^d$, there exists a positive constant C such that

$$\left| \mathbb{P}\left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \in S \right) - \mathbb{P}(Z \in S) \right| \leq C\, d^{1/4} n^{-1/2}\, \mathbf{E}[\, \| \Sigma^{-1/2}\, X_1 \|_2^3\,].$$

Proof of Proposition 8.
For the trust region data set T_k, let

$$U^k \triangleq \begin{pmatrix} (x^1)^\top \\ \vdots \\ (x^{N_k})^\top \end{pmatrix}, \quad \widehat{U}^k \triangleq \begin{pmatrix} 1 & (x^1)^\top \\ \vdots & \vdots \\ 1 & (x^{N_k})^\top \end{pmatrix}, \quad W^k \triangleq \begin{pmatrix} (\omega^1)^\top \\ \vdots \\ (\omega^{N_k})^\top \end{pmatrix}, \quad \psi(U^k) \triangleq \begin{pmatrix} \psi(x^1)^\top \\ \vdots \\ \psi(x^{N_k})^\top \end{pmatrix}.$$

The matrix of error is defined to be $\Xi^k \triangleq W^k - \psi(X^k)$. Under the poisedness condition, least square estimators of the LLR model are unique and can be written in matrix form.

$$\begin{pmatrix} \widehat{B}^{k,0} \\ \widehat{B}^{k,1} \end{pmatrix} = \left((\widehat{U}^k)^\top \widehat{U}^k \right)^{-1} (\widehat{U}^k)^\top\, W^k \tag{25}$$

Under $(A1)$ and $(A3)$, by the Taylor's expansion of ψ at $z \in B(\widehat{x}^k, \delta_k) \cap X$, we derive

$$\psi(x^i) = \psi(z) + \nabla\psi(z)^\top (x^i - z) + O(L_2\, \delta_k^2)\ \mathbf{1}_m, \qquad \forall i = 1, \ldots, |T_k|.$$

We rewrite the Taylor's expansion in the matrix form.

$$\begin{aligned} \psi(U^k) &= \mathbf{1}_{N_k} \psi(z)^\top + (U^k - \mathbf{1}_{N_k} z^\top)\, \nabla\psi(z) + O(L_2\, \delta_k^2)\ \mathbf{1}_{N_k \times m} \\ &= \widehat{U}^k \begin{pmatrix} \psi(z)^\top - z^\top \nabla\psi(z) \\ \nabla\psi(z) \end{pmatrix} + O(L_2\, \delta_k^2)\ \mathbf{1}_{N_k \times m}. \end{aligned}$$

Under the poisedness condition, we derive

$$\begin{pmatrix} \psi(z)^\top - z^\top \nabla\psi(z) \\ \nabla\psi(z) \end{pmatrix} = \left((\widehat{U}^k)^\top \widehat{U}^k \right)^{-1} \left(\widehat{U}^k \right)^\top \left(\psi(U^k) + O(L_2\, \delta_k^2)\ \mathbf{1}_{N_k \times m} \right) \tag{26}$$

By subtracting (26) from the LLR estimation (25), we derive

$$\begin{pmatrix} \widehat{B}^{k,0} \\ \widehat{B}^{k,1} \end{pmatrix} - \begin{pmatrix} \psi(z)^\top - z^\top \nabla\psi(z) \\ \nabla\psi(z) \end{pmatrix} = \left((\widehat{U}^k)^\top \widehat{U}^k\right)^{-1} \left(\widehat{U}^k\right)^\top \left(\Xi^k + O(L_2\,\delta_k^2)\;\mathbf{1}_{N_k\times m}\right) \tag{27}$$

The above equation (27) provides the estimates for the gap $\nabla\psi(z) - \widehat{B}^{k,1}$ and $\psi(z) - (\widehat{B}^{k,0})^\top - \nabla\psi(z)^\top z$. Next we consider the asymptotical convergence of two LLR estimators, $\nabla\psi(z) - \widehat{B}^{k,1}$ and $\psi(z) - (\widehat{B}^{k,0})^\top - (\widehat{B}^{k,1})^\top z$ on $\mathcal{B}(\widehat{x}^k, \delta_k) \cap X$ respectively.

(a) Let $\bar{x} = \sum_{i=1}^{N_k} x^i/N_k$, and $\bar{U}^k \triangleq (\mathbb{I}_{N_k} - \frac{1}{N_k}\mathbf{1}_{N_k\times N_k})\,U^k = \begin{pmatrix} \left(x^1 - \bar{x}\right)^\top \\ \vdots \\ \left(x^{N_k} - \bar{x}\right)^\top \end{pmatrix}$. Let $\widehat{\mathbb{I}}_p \triangleq (\mathbf{0}_{p\times 1}, \mathbb{I}_p)$. From the inverse of block matrix $(\widehat{U}^k)^\top \widehat{U}^k$, we derive that,

$$\widehat{\mathbb{I}}_p \left((\widehat{U}^k)^\top \widehat{U}^k\right)^{-1} (\widehat{U}^k)^\top = (\mathbf{0}_{p\times 1}, \mathbb{I}_p) \begin{pmatrix} * & * \\ C_k & D_k \end{pmatrix} \begin{pmatrix} \mathbf{1}_{N_k}^\top \\ (\widehat{U}^k)^\top \end{pmatrix} = C_k\,\mathbf{1}_{N_k}^\top + D_k\,(\widehat{U}^k)^\top = \left((\bar{U}^k)^\top \bar{U}^k\right)^{-1} (\bar{U}^k)^\top$$

where $*$ represents the matrices on the first row of the matrix, and

$$\begin{aligned} C_k &\triangleq -\frac{1}{N_k}\,D_k\,(\widehat{U}^k)^\top \mathbf{1}_{N_k} = -\frac{1}{N_k}\left((\bar{U}^k)^\top \bar{U}^k\right)^{-1} (\widehat{U}^k)^\top \mathbf{1}_{N_k} \\ D_k &\triangleq \left[(\widehat{U}^k)^\top \widehat{U}^k - \frac{1}{N_k}(\widehat{U}^k)^\top \mathbf{1}_{N_k}\mathbf{1}_{N_k}^\top \widehat{U}^k\right]^{-1} = ((\bar{U}^k)^\top \bar{U}^k)^{-1}. \end{aligned}$$

By (27) we derive

$$\begin{aligned} \nabla\psi(z) - \widehat{B}_1^k &= \widehat{\mathbb{I}}_p \left((\widehat{U}^k)^\top \widehat{U}^k\right)^{-1} (\widehat{U}^k)^\top \Xi^k + O(L_2\,\delta_k^2)\;\widehat{\mathbb{I}}_p \left((\widehat{U}^k)^\top \widehat{U}^k\right)^{-1} (\widehat{U}^k)^\top \mathbf{1}_{N_k\times m} \\ &= \left((\bar{U}^k)^\top \bar{U}^k\right)^{-1} (\bar{U}^k)^\top \Xi^k + O(L_2\,\delta_k^2)\;\left((\bar{U}^k)^\top \bar{U}^k\right)^{-1} (\bar{U}^k)^\top \mathbf{1}_{N_k\times m} \\ &= \frac{1}{\delta_k}\left((\widetilde{U}^k)^\top \widetilde{U}^k\right)^{-1} (\widetilde{U}^k)^\top \Xi^k \end{aligned}$$

where $\widetilde{U}^k \triangleq \bar{U}^k/\delta_k$. According to the construction of $\{x^i\}$ at line 3 in CLEO algorithm,

$$\widetilde{U}^k = \begin{pmatrix} (\widetilde{u}^1 - \bar{u})^\top \\ \vdots \\ (\widetilde{u}^{N_k} - \bar{u})^\top \end{pmatrix}$$

where each $\widetilde{u}^i$ is independently sampled from $\mathcal{U}(\mathcal{B}(\mathbf{0}_p, 1))$ for $i = 1, \ldots, N_k$ and $\bar{u} = \frac{1}{N_k}\sum_{i=1}^{N_k} u^i$. Now we show that the matrix $\left((\widetilde{U}^k)^\top \widetilde{U}^k\right)^{-1} (\widetilde{U}^k)^\top \Xi^k$ asymptotically converges in distribution. By the Law of Large Numbers, $\frac{1}{N_k}(\widetilde{U}^k)^\top \widetilde{U}^k \to V_u$ almost surely where $V_u \triangleq \mathbf{E}_{\widetilde{x}}\left[\widetilde{u}\,\widetilde{u}^\top\right]$ and $\widetilde{u} \sim \mathcal{U}(\mathcal{B}(\mathbf{0}_p, 1))$. By Theorem 2.57 in [17], $\frac{1}{N_k}(\widetilde{U}^k)^\top \widetilde{U}^k$ converges with the rate of $o(N_k^{-1/2})$. By the Berry-Esseen Theorem, there exists a constant C such that with the rate of $CN_k^{-1/2}$, we have

$$\frac{1}{\sqrt{N_k}}\,(\widetilde{U}^k)^\top \Xi^k \xrightarrow{d} \mathcal{MN}_{p\times m}(0, V_u, \Sigma), \qquad \text{as } N_k \to \infty.$$

By the Slutsky's Theorem, with the rate of $O(N_k^{-1/2})$,

$$\sqrt{N_k}\left((\widetilde{U}^k)^\top\widetilde{U}^k\right)^{-1}(\widetilde{U}^k)^\top\Xi^k \xrightarrow{d} \mathcal{MN}_{p\times m}(0, V_u^{-1}, \Sigma), \qquad \text{as } N_k \to \infty. \tag{28}$$

Let $H_1 \sim \mathcal{MN}_{p\times m}(0, V_u^{-1}, \Sigma)$ be a random matrix and let Ψ_1 be a cumulative probability distribution of $\|H_1\|$. For any $0 < \alpha < 1$, there exists a constant $\kappa_{eg} > 0$, when $N_k \geq \max\left\{\delta_k^{-4}\left(\Psi_1(1-\alpha/2)\right)^2\kappa_{eg}^{-2}, 4C^2\alpha^{-2}\right.$ we have

$$\begin{aligned}
&\mathbb{P}\left(\|\nabla\psi(z) - \widehat{B}^{k,1}\| \geq \kappa_{eg}\Delta_k,\ \forall z \in \mathcal{B}(\widehat{X}^k, \Delta_k)\cap X \,\middle|\, \mathcal{F}_{k-1}\right)\\
\leq&\mathbb{P}\left(\|(\Delta_k)^{-1}\left((\widetilde{U}^k)^\top\widetilde{U}^k\right)^{-1}(\widetilde{U}^k)^\top\Xi^k\| \geq \kappa_{eg}\Delta_k \mid \mathcal{F}_{k-1}\right)\\
\leq&\mathbb{P}\left(\frac{1}{\sqrt{N_k}}\|H_1\| \geq \kappa_{eg}(\Delta_k)^2 \,\middle|\, \mathcal{F}_{k-1}\right) + CN_k^{-1/2}\\
\leq&\alpha/2 + \alpha/2 = \alpha.
\end{aligned}$$

(b) Consider

$$\begin{aligned}
&\psi(z) - (\widehat{B}^{k,1})^\top z - (\widehat{B}^{k,0})^\top\\
=&\left[\psi(z) - \nabla\psi(z)^\top z - (\widehat{B}^{k,0})^\top\right] + (\nabla\psi(z) - \widehat{B}^{k,1})^\top z\\
=&\begin{pmatrix}\psi(z)^\top - z^\top\nabla\psi(z) - \widehat{B}^{k,0}\\ \nabla\psi(z) - \widehat{B}^{k,1}\end{pmatrix}^\top\begin{pmatrix}1\\ z\end{pmatrix}\\
=&-\underbrace{(\Xi^k)^\top(\widehat{U}^k)\left((\widehat{U}^k)^\top\widehat{U}^k\right)^{-1}\begin{pmatrix}1\\ z\end{pmatrix}}_{T1} + O(\delta_k^2)\ \underbrace{\mathbf{1}_{m\times N_k}(\widehat{U}^k)\left((\widehat{U}^k)^\top\widehat{U}^k\right)^{-1}\begin{pmatrix}1\\ z\end{pmatrix}}_{T2}
\end{aligned} \tag{29}$$

By definition 7, $l(z) = (\widehat{U}^k)((\widehat{U}^k)^\top\widehat{U}^k)^{-1}\begin{pmatrix}1\\ z\end{pmatrix}$. Under the strong poisedness condition, we have $\sqrt{N_k}\ \max_{z\in\mathcal{B}}\|\,l(z)\,\| \leq \Lambda$. Therefore,

$$\begin{aligned}
\mathbf{E}\left[\|\,T_1\,\|^2\right] &= \mathbf{E}\left[\,\|\,(\Xi^k)^\top l(z)\,\|^2\,\right] \leq \textstyle\sum_{i=1}^{N_k}\mathbf{E}\left[\|\,\varepsilon^i\,\|^2\right]\cdot\mathbf{E}\left[\,|\,l_i(z)\,|^2\right] \leq \frac{\Lambda^2}{N_k}\mathbf{E}[\,\|\,\widetilde{\varepsilon}\,\|^2]\\
\|\,T_2\,\| &= \|\,\mathbf{1}_{m\times N_k}l(z)\,\| = \sqrt{m}\,|\mathbf{1}_{N_k}^\top l(x)| \leq \sqrt{N_k}\,\sqrt{m}\,\|\,l(x)\,\| \ \leq\ \Lambda\sqrt{m}.
\end{aligned}$$

From the Markov's Inequality, there exists a constant C such that

$$\begin{aligned}
&\mathbb{P}\left(\|\,\psi(z) - (\widehat{B}^{k,1})^\top z - (\widehat{B}^{k,0})^\top\,\| \geq \kappa_{ef}\Delta_k^2,\ \forall z \in \mathcal{B}(\widehat{X}^k, \Delta_k)\cap X \,\middle|\, \mathcal{F}_{k-1}\right)\\
\leq\ &\mathbb{P}\left(\|T_1\| + C\,\Delta_k^2\|T_2\| \geq \kappa_{ef}\Delta_k^2,\ \forall z \in \mathcal{B}(\widehat{X}^k, \Delta_k)\cap X \,\middle|\, \mathcal{F}_{k-1}\right)\\
\leq\ &\mathbb{P}\left(\|T_1\| \geq (\kappa_{ef} - C\Lambda\sqrt{m})\Delta_k^2,\ \forall z \in \mathcal{B}(\widehat{X}^k, \Delta_k)\cap X \,\middle|\, \mathcal{F}_{k-1}\right)\\
\leq\ &\frac{\mathbf{E}\left[\|T_1\|^2\right]}{(\kappa_{ef} - C\Lambda\sqrt{m})^2\,\delta_k^4} \ \leq\ \frac{\Lambda^2\mathbf{E}[\,\|\,\widetilde{\varepsilon}\,\|^2]}{N_k(\kappa_{ef} - C\Lambda\sqrt{m})^2\,\delta_k^4}.
\end{aligned}$$

Therefore for any $0 < \alpha < 1$, there exists a constant $\kappa_{ef} > 0$, when $N_k \geq O(\delta_k^{-4}\kappa_{ef}^{-2}\alpha^{-1})$, the statement (b) holds. □

We need the following lemma for the proof of Proposition 9.

Lemma 17. With a finite index set $\mathcal{J}$, for each $j \in \mathcal{J}$, suppose $h_j(\cdot) : Z \subseteq \mathbb{R}^n \to \mathbb{R}$ is a smooth convex function on a compact set Z. Let $h(z) \triangleq \max_{j\in\mathcal{J}} h_j(z)$. For any $z \in Z$, $\mathcal{A}(z) \triangleq \{j \in \mathcal{J} : h_j(z) \geq h(z)\}$, $\mathcal{A}_\epsilon(z) \triangleq \{j \in \mathcal{J} : h_j(z) \geq h(z) - \epsilon\}$ for any $\epsilon > 0$, and $L_1 \triangleq \max\{ \| \nabla h_j(z) \| : j \in \mathcal{J}, z \in Z \}$. For any $z \in Z$, any $\delta \geq 0$, any $z' \in \mathcal{B}(z, \delta)$ and any $\epsilon \geq 2L_1\delta$, we have $\mathcal{A}(z') \subseteq \mathcal{A}_\epsilon(z)$.

Proof. For the max of convex functions, this lemma establishes the relation between the active index set $\mathcal{A}(z')$ and ϵ-active index set $\mathcal{A}_\epsilon(z)$ for any $z' \subseteq \mathcal{B}(z, \delta)$ with an appropriate choice of ϵ.

Specifically, for any $z \in Z$, any $\delta \geq 0$ and any $z' \in \mathcal{B}(z, \delta)$, we have for each $j \in \mathcal{A}(z')$,

$$h_j(z) = h(z) - h(z) + h(z') - h(z') + h_j(z') - h_j(z') + h_j(z) \geq h(z) - 2L\delta.$$

Hence $j \in \mathcal{A}_\varepsilon(z)$ and $\mathcal{A}(z') \subseteq \mathcal{A}_\varepsilon(z)$. □

Proof of Proposition 9.
For each (x^i, ω^i), let $\varepsilon^i \triangleq \omega^i - \psi(x^i)$ and the residual of LLR model is $e^{k,i} \triangleq \omega^i - (\widehat{B}^{k,1})^\top x^i - (\widehat{B}^{k,0})^\top = \psi(x^i) - (\widehat{B}^{k,1})^\top x^i - (\widehat{B}^{k,0})^\top + \varepsilon^i$. For any $z \in \mathcal{B}(\widehat{x}^k, \delta_k) \cap X$,

$$\begin{aligned}
\left| f(z) - f_k(z) \right| &\leq \left| \mathbf{E}_{\widetilde{\varepsilon}}[\, h(z, \psi(z) + \widetilde{\varepsilon})\,] - \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} h(z, \psi(z) + \varepsilon^i) \right| \\
&\quad + \left| \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} h(z, \psi(z) + \varepsilon^i) - \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} h(z, m_k(z, e^{k,i})) \right| \\
&\quad + \left| \mathbf{E}_{\widetilde{\varepsilon}}[\, g(z, \psi(z) + \widetilde{\varepsilon})\,] - \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} g(z, \psi(z) + \varepsilon^i) \right| \\
&\quad + \left| \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} g(z, \psi(z) + \varepsilon^i) - \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} g(z, m_k(z, e^{k,i})) \right| \\
&\leq |\tau_k(z)| + |\xi_k(z)| + 2L_1 \max_{z\in\mathcal{B}(\widehat{X}^k, \Delta_k)\cap X} \| \psi(z) - (\widehat{B}^{k,1})^\top z - (\widehat{B}^{k,0})^\top \|
\end{aligned}$$

where

$$\begin{aligned}
\tau_k(z) &\triangleq \mathbf{E}_{\widetilde{\varepsilon}}[\, h(z, \psi(z) + \widetilde{\varepsilon})\,] - \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} h(z, \psi(z) + \varepsilon^i), \\
\xi_k(z) &\triangleq \mathbf{E}_{\widetilde{\varepsilon}}[\, g(z, \psi(z) + \widetilde{\varepsilon})\,] - \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} g(z, \psi(z) + \varepsilon^i),
\end{aligned}$$

and L_1 is the Lipschitz continuous modulus for functions $h(z, \cdot)$ and $g(z, \cdot)$.
Consider the stochastic process F_k given $\mathcal{F}_{k-1}$.

$$\begin{aligned}
&\mathbb{P}\left(|f(z) - f_k(z)| \geq \kappa_{ef}\Delta_k^2, \forall z \in \mathcal{B}(\widehat{X}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1} \right) \\
\leq\ &\mathbb{P}\left(|\tau_k(z)| \geq \frac{1}{3}\kappa_{ef}\Delta_k^2, \forall z \in \mathcal{B}(\widehat{X}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1} \right) \\
&+ \mathbb{P}\left(|\xi_k(z)| \geq \frac{1}{3}\kappa_{ef}\Delta_k^2, \forall z \in \mathcal{B}(\widehat{X}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1} \right) \\
&+ \mathbb{P}\left(2L_1 \max_{z\in\mathcal{B}(\widehat{X}^k, \Delta_k)\cap X} \| \psi(z) - (\widehat{B}^{k,1})^\top z - (\widehat{B}^{k,0})^\top \| \geq \frac{1}{3}\kappa_{ef}\Delta_k^2 \mid \mathcal{F}_{k-1} \right) \qquad (30)
\end{aligned}$$

Since $\{\varepsilon^i\}$ are iid samples of the bounded random variable $\widetilde{\varepsilon}$, the Berry-Esseen Theorem implies that with the rate $O(1/\sqrt{N_k})$, we have $\sqrt{N_k}\,\tau_k(z) \stackrel{d}{\to} \mathcal{N}(0, V_h(z))$ where $V_h(z) = Var_{\widetilde{\varepsilon}}(h(z, \psi(z) + \widetilde{\varepsilon}))$. Similarly, with rate $O(1/\sqrt{N_k})$, we have $\sqrt{N_k}\,\xi_k(z) \stackrel{d}{\to} \mathcal{N}(0, V_g(z))$ where $V_g(z) = Var_{\widetilde{\varepsilon}}(g(z, \psi(z)+\widetilde{\varepsilon}))$. Moreover, with Proposition 8, there exists a constant $\kappa_{ef} > 0$, when $N_k \geq \max\{O(\delta_k^{-4}\kappa_{ef}^{-4}),\, O(\alpha^{-2})\}$, each term on the right-hand side of (30) is bounded by $\alpha/6$. Therefore, we establish part (a).

For part (b), we denote $\partial_\gamma h(x,\omega)$ the γ-subdifferential of the convex function h for any $\gamma > 0$, i.e.,

$$\partial_\gamma h(x,\omega) \triangleq \{(q_1, q_2) \in \mathbb{R}^{m+p} : h(x',\omega') \geq h(x,\omega) + q_1^\top (x' - x) + q_2^\top (\omega' - \omega) - \gamma\}.$$

The γ-directional derivative of $h(x,\omega)$ with the direction $(d_1, d_2) \in \mathbb{R}^{m+p}$ is defined as

$$h'_\gamma((x,\omega);(d_1,d_2)) \triangleq \max\{\, d_1^\top q_1 + d_2^\top q_2 : (q_1,q_2) \in \partial_\gamma h(x,\omega)\,\}.$$

Moreover, we denote an γ-approximate directional derivative based on the γ-subdifferential by $\widehat{\chi}_\epsilon(x)$.

$$\widehat{\chi}_\gamma(x) \triangleq \Big|\min_{\substack{x+d\in X\\ \|d\|\leq 1}} \left(\begin{array}{l} \nabla c(x)^\top d + \mathbf{E}_{\widetilde{\varepsilon}}\,\big[\,h'_\gamma((x,\psi(x)+\widetilde{\varepsilon});(d,\nabla\psi(x)^\top d))\big] \\ +\mathbf{E}_{\widetilde{\varepsilon}}[\,g'((x,\psi(x)+\widetilde{\varepsilon});(d,\nabla\psi(x)^\top d))\,], \end{array}\right)\Big|. \tag{31}$$

Let $\kappa_L \triangleq 2\kappa_{ef}\cdot\max\{\,\|\nabla h_j(x,\omega)\| : j\in\mathcal{J}, (x,\omega)\in X\times\Omega\,\}$ and $\bar{\epsilon}_k \triangleq \kappa_L\delta_k^2$. For any $x \in \mathcal{B}(\widehat{x}^k,\delta_k)\cap X$, let $\bar{d}_k$ be the optimal direction in defining $\widehat{\chi}_{\bar{\epsilon}_k}(x)$. Then

$$\begin{aligned}
\chi(x) - \chi_k(x) &\leq \widehat{\chi}_{\bar{\epsilon}_k}(x) - \chi_k(x) \\
&\leq f_k'(x;\,\bar{d}_k) - \left(\begin{array}{l} \nabla c(x)^\top \bar{d}_k + \mathbf{E}_{\widetilde{\varepsilon}}\,\big[\,h'_{\bar{\epsilon}_k}\big((x,\psi(x)+\widetilde{\varepsilon});(\bar{d}_k,\nabla\psi(x)^\top\bar{d}_k)\big)\big] \\ +\mathbf{E}_{\widetilde{\varepsilon}}[\,g'\,((x,\psi(x)+\widetilde{\varepsilon});(\bar{d}_k,\,\nabla\psi(x)^\top\,\bar{d}_k))\,], \end{array}\right) \\
&= \frac{1}{N_k}\sum_{i=1}^{N_k}\Big(\,h'((x,m_k(x,e^{\,k,i});(\bar{d}_k,(\widehat{B}^{k,1})^\top\bar{d}_k)) + g'((x,m_k(x,e^{\,k,i}));(\bar{d}_k,(\widehat{B}^{k,1})^\top\bar{d}_k))\Big) \\
&\quad -\mathbf{E}_{\widetilde{\varepsilon}}\,\big[\,h'_{\bar{\epsilon}_k}\,\big((x,\psi(x)+\widetilde{\varepsilon});(\bar{d}_k,\nabla\psi(x)^\top\bar{d}_k)\big) + g'\,((x,\psi(x)+\widetilde{\varepsilon});(\bar{d}_k,\,\nabla\psi(x)^\top\,\bar{d}_k))\,\big] \\
&= \eta_{k1}(x) + \eta_{k2}(x) + \gamma_{k1}(x) + \gamma_{k2}(x),
\end{aligned}$$

where

$$\eta_{k1}(x) \triangleq -\mathbf{E}_{\widetilde{\varepsilon}}\Big[h'_{\bar{\epsilon}_k}\,\Big((x,\psi(x)+\widetilde{\varepsilon});(\bar{d}_k,\nabla\psi(x)^\top\bar{d}_k)\Big)\Big] + \frac{1}{N_k}\sum_{i=1}^{N_k} h'_{\bar{\epsilon}_k}\,\Big((x,\psi(x)+\varepsilon^i);(\bar{d}_k,\nabla\psi(x)^\top\bar{d}_k)\Big),$$

$$\eta_{k2}(x) \triangleq \frac{1}{N_k}\sum_{i=1}^{N_k} h'((x,m_k(x,e^{k,i}));(\bar{d}_k,(\widehat{B}^{k,1})^\top\,\bar{d}_k)) - \frac{1}{N_k}\sum_{i=1}^{N_k} h'_{\bar{\epsilon}_k}\,\Big((x,\psi(x)+\varepsilon^i);(\bar{d}_k,\nabla\psi(x)^\top\bar{d}_k)\Big),$$

$$\gamma_{k1}(x) \triangleq -\mathbf{E}_{\widetilde{\varepsilon}}\Big[g'\,((x,\psi(x)+\widetilde{\varepsilon});(\bar{d}_k,\,\nabla\psi(x)^\top\,\bar{d}_k))\Big] + \frac{1}{N_k}\sum_{i=1}^{N_k} g'\,((x,\psi(x)+\varepsilon^i);(\bar{d}_k,\,\nabla\psi(x)^\top\,\bar{d}_k)),$$

$$\gamma_{k2}(x) \triangleq \frac{1}{N_k}\sum_{i=1}^{N_k} g'((x,m_k(x,e^{\,k,i}));(\bar{d}_k,(\widehat{B}^{k,1})^\top\bar{d}_k)) - \frac{1}{N_k}\sum_{i=1}^{N_k} g'\,((x,\psi(x)+\varepsilon^i);(\bar{d}_k,\,\nabla\psi(x)^\top\,\bar{d}_k)).$$

By the Berry-Esseen Theorem, since $\widetilde{\varepsilon}$ is a bounded random variable, we can derive that $\sqrt{N_k}\eta_{k1}(x)$ and $\sqrt{N_k}\gamma_{k1}(x)$ converge in distribution to a normal distribution with the rate $O(1/\sqrt{N_k})$ uniformly for all $x \in \mathcal{B}(\widehat{x}^k, \delta_k) \cap X$. We then bound $\eta_{k2}(x)$. From Proposition 8, for any $0 < \alpha < 1$, for any $N_k \geq \max\{O(\delta_k^{-4}\kappa_{eg}^{-2}),\, O(\alpha^{-2}), O(\delta_k^{-4}\kappa_{ef}^{-2}\alpha^{-1})\}$, we have

$$\mathbb{P}\left(\begin{array}{l} \left\| \psi(x) - (\widehat{B}^{k,1})^\top x - (\widehat{B}^{k,0})^\top \right\| \leq \kappa_{ef}\, \Delta_k^2, \\ \left\| \nabla\psi(x) - \widehat{B}^{k,1} \right\| \leq \kappa_{eg}\Delta_k, \forall x \in \mathcal{B}(\widehat{X}^k, \Delta_k) \cap X \end{array} \Big| \, \mathcal{F}_k \right) \geq 1 - \alpha/8. \tag{32}$$

Based on the definition, we know that

$$h'_{\bar{\epsilon}_k}\left((x, \psi(x) + \varepsilon^i); (\bar{d}_k, \nabla\psi(x)^\top \bar{d}_k)\right) = \max\left\{ \bar{d}_k^\top q_1 + \bar{d}_k^\top \nabla\psi(x)\, q_2 : (q_1, q_2) \in \partial_{\bar{\epsilon}_k} h(x, \psi(x) + \varepsilon^i) \right\},$$

$$h'((x, m_k(x, e^{k,i}));(\bar{d}_k, (\widehat{B}^{k,1})^\top \bar{d}_k)) = \max\left\{ \bar{d}_k^\top q_1 + \bar{d}_k^\top \widehat{B}^{k,1}\, q_2 : (q_1, q_2) \in \partial\, h(x, m_k(x, e^{k,i})) \right\}.$$

For the max of convex functions, it is known that $\partial h(x,\omega) = \mathrm{conv}\{\nabla h_j(x,\omega) : j \in \mathcal{A}(x,\omega)\}$ and $\mathrm{conv}\{\nabla h_j(x,\omega) : j \in \mathcal{A}_\epsilon(x,\omega)\} \subseteq \partial_\epsilon h(x,\omega)$ for any $\epsilon > 0$. Since

$$\left\| \psi(x) + \varepsilon^i - m_k(x, e^{k,i}) \right\| \leq \left\| \psi(x) - (\widehat{B}^{k,1})^\top x - (\widehat{B}^{k,0})^\top \right\| + \left\| \psi(x^i) - (\widehat{B}^{k,1})^\top x^i - (\widehat{B}^{k,0})^\top \right\|,$$

from Lemma 5, under the condition that the event in (32) holds, for any $x \in \mathcal{B}(\widehat{x}^k, \delta_k) \cap X$,

$$\mathcal{A}(x, m_k(x, e^{k,i})) \subseteq \mathcal{A}_{\bar{\epsilon}_k}(x, \psi(x) + \varepsilon^i) \quad \forall i = 1, \ldots, N_k.$$

Hence, under the same condition, for any $x \in \mathcal{B}(\widehat{x}^k, \delta_k) \cap X$, we can derive that

$$\begin{aligned} \eta_{k2}(x) \;&\leq \frac{1}{N_k}\sum_{i=1}^{N_k} \max_{j \in \mathcal{A}(x, m_k(x, e^{k,i}))} \left\{ \bar{d}_k^\top \nabla_1 h_j(x, m_k(x, e^{k,i})) + \bar{d}_k^\top \widehat{B}^{k,1}\, \nabla_2 h_j(x, m_k(x, e^{k,i})) \right\} \\ &\quad - \frac{1}{N_k}\sum_{i=1}^{N_k} \max_{j \in \mathcal{A}_{\bar{\epsilon}_k}(x, \psi(x) + \varepsilon^i)} \left\{ \bar{d}_k^\top \nabla_1 h_j(x, \psi(x) + \varepsilon^i) + \bar{d}_k^\top \nabla\psi(x)\, \nabla_2 h_j(x, \psi(x) + \varepsilon^i) \right\} \\ &\leq \frac{1}{N_k}\sum_{i=1}^{N_k} \max_{j \in \mathcal{J}} \left\| \nabla_1 h_j(x, m_k(x, e^{k,i})) - \nabla_1 h_j(x, \psi(x) + \varepsilon^i) \right\| \\ &\quad + \frac{1}{N_k}\sum_{i=1}^{N_k} \max_{j \in \mathcal{J}} \left\| (\widehat{B}^{k,1})\, \nabla_2 h_j(x, m_k(x, e^{k,i})) - \nabla\psi(x)\, \nabla_2 h_j(x, \psi(x) + \varepsilon^i) \right\| \\ &\leq 2\,(L_1 L_2 + L_2) \max_{x \in \mathcal{B}(\widehat{x}^k, \delta_k) \cap X} \left\| \psi(x) - (\widehat{B}^{k,1})^\top x - (\widehat{B}^{k,0})^\top \right\| + L_1 \| \nabla\psi(x) - \widehat{B}^{k,1} \|. \end{aligned}$$

For the term $\gamma_{k2}(x)$, we derive for any $x \in \mathcal{B}(\widehat{x}^k, \delta_k) \cap X$,

$$\begin{aligned} \gamma_{k2}(x) \;&=\; \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} \left[-g'\left((x, \psi(x) + \varepsilon^i); (\bar{d}_k,\, \nabla\psi(x)^\top \bar{d}_k)\right) + g'((x, m_k(x, e^{k,i}); (\bar{d}_k,\, \nabla\psi(x)^\top \bar{d}_k)) \right] \\ &\quad + \tfrac{1}{N_k}\textstyle\sum_{i=1}^{N_k} \left[-g'((x, m_k(x, e^{k,i}); (\bar{d}_k,\, \nabla\psi(x)^\top \bar{d}_k)) + g'((x, m_k(x, e^{k,i})); (\bar{d}_k,\, (\widehat{B}^{k,1})^\top \bar{d}_k)) \right] \\ &\leq\; 2(L_2 + L_1 L_2) \max_{x \in \mathcal{B}(\widehat{x}^k, \delta_k) \cap X} \left\| \psi(x) - (\widehat{B}^{k,1})^\top x - (\widehat{B}^{k,0})^\top \right\| + L_1 \left\| \nabla\psi(x) - \widehat{B}^{k,1} \right\|, \end{aligned}$$

where L_1 is the Lipschitz continuity modulus and L_2 is the Lipschitz gradient modulus of the function g. Now we can combine the bounds of $\eta_{k,1}$, $\eta_{k,2}$, $\gamma_{k,1}$, and $\gamma_{k,2}$. Consider the stochastic

process F_k given $\mathcal{F}_{k-1}$. With Proposition 8, there exists a constant $\kappa_{ed} > 0$, such that for any $N_k \geq \max\{O(\delta_k^{-4}\kappa_{ed}^{-2}\alpha^{-1}),\, O(\alpha^{-2})\}$, we have

$$\begin{aligned}
& \mathbb{P}\left(\chi(x) - \chi_k(x) \geq \kappa_{ed}\Delta_k, \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1}\right) \\
\leq\ & \mathbb{P}\left(\widehat{\chi}_{\bar{\epsilon}_k}(x) - \chi_k(x) \geq \kappa_{ed}\Delta_k, \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1}\right) \\
\leq\ & \mathbb{P}\left(\eta_{k1}(x) + \eta_{k2}(x) + \gamma_{k1}(x) + \gamma_{k2}(x) \geq \kappa_{ed}\Delta_k, \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1}\right) \\
\leq\ & \mathbb{P}\left(\eta_{k1}(x) \geq \tfrac{1}{4}\kappa_{ed}\Delta_k, \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1}\right) + \mathbb{P}\left(\eta_{k2}(x) \geq \tfrac{1}{4}\kappa_{ed}\Delta_k, \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1}\right) \\
& + \mathbb{P}\left(\gamma_{k1}(x) \geq \tfrac{1}{4}\kappa_{ed}\Delta_k, \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1}\right) + \mathbb{P}\left(\gamma_{k2}(x) \geq \tfrac{1}{4}\kappa_{ed}\Delta_k, \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \mid \mathcal{F}_{k-1}\right) \\
\leq\ & 4 \cdot \tfrac{\alpha}{8} = \tfrac{\alpha}{2},
\end{aligned}$$

which establishes part (b).

Combining the results in (a) and (b), we have for any $0 < \alpha < 1$, there exists a constant vector $\kappa = (\kappa_{ef}, \kappa_{ed})$, for any $N_k \geq \max\{O(\delta_k^{-4}\kappa_{ed}^{-2}\alpha^{-1}),\, O(\delta_k^{-4}\kappa_{ef}^{-2}),\, O(\alpha^{-2})\}$ we have

$$\mathbb{P}\left(|\, f(x) - F_k(x)\,| \,\leq\, \kappa_{ef}\,\Delta_k^2,\ \chi(x) - \chi_k(x) \,\leq\, \kappa_{ed}\,\Delta_k,\ \forall x \in \mathcal{B}(\widehat{x}^k, \Delta_k) \cap X \,\middle|\, \mathcal{F}_{k-1}\right) \geq 1 - \alpha,$$

which establishes the statement. □

Proof of Proposition 14. The proof of this result is similar to Theorem 4.11 in [9] except that $\|\nabla f(\widehat{x}_k)\|$ in the last reference for smooth functions need to be replaced by $\chi(\widehat{x}^k)$, the nonsmooth counterpart. Due to this change, we present the proof with several modifications here.

Define the random function $\Phi_k = \nu f(\widehat{X}_k) + (1-\nu)(\Delta_k^2 - \Delta_k)$ where $\nu \in (0,1)$ satisfying

$$\frac{\nu}{1-\nu} > \max\left\{\frac{4\gamma^2}{\zeta C_1}, \frac{4\gamma^2}{C_2}, \frac{\gamma^2}{C_3}\right\}, \tag{33}$$

where C_1 and C_2 are defined in Lemma 11 and Lemma 13. The goal of the analysis is to show that there exists constants $\tau > 0$ such that for all k

$$\mathbb{E}\left[\Phi_{k+1} - \Phi_k \mid \mathcal{F}_{k-1}\right] \leq -\tau\Delta_k^2 < 0. \tag{34}$$

Since f is bounded from below by assumption, and $\Delta_k \in [0, \delta_{\max}]$, Φ_k is bounded from below for all k. By summing over the above inequality over k and taking expectations on both sides, we can conclude that $\sum_{k=0}^{\infty}\Delta_k$ is finite almost surely. To show (34) holds for every iteration, we discuss 2 cases depending on whether $\chi(\widehat{x}^k) \geq \zeta\delta_k$ or not, where

$$\zeta \geq\ \delta_{\max}\,\kappa_{ed} + \max\left\{\eta_2,\ \frac{4\kappa_{ef}\,\delta_{\max}}{\kappa_{dcp}(1-\eta_1)}\right\}.$$

Within each of these cases, there are 4 combined outcomes of the event I_k and J_k as defined in Definition 3 and 5 with different probabilities. For each outcome we develop an upper bound of the difference $\phi_{k+1} - \phi_k$ where ϕ_k denotes the outcome of Φ_k. Then we bound $\mathbb{E}[\Phi_{k+1} - \Phi_k \mid \mathcal{F}_{k-1}]$ by combining the bounds of 4 outcomes, which results in (34) for all iterations. We only analyze one case that $\chi(\widehat{x}^k) \geq \zeta\delta_k$ to obtain (34), since the analysis of the other case can be completed with similar modifications based on the proof of Theorem 4.11 in [9].

We first analyze $\phi_{k+1} - \phi_k$ for successful and unsuccessful iterations respectively. For all successful iterations, $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \gamma\,\delta_k$ with $\gamma > 1$, hence,

$$\phi_{k+1} - \phi_k \;=\; \nu(f(\widehat{x}^{k+1}) - f(\widehat{x}^k)) + (1-\nu)(\gamma^2 - 1)\delta_k^2 + (1-\nu)(1-\gamma)\delta_k. \tag{35}$$

For all unsuccessful iterations, $x_{k+1} = x_k$ and $\delta_{k+1} = \delta_k/\gamma$, hence

$$\phi_{k+1} - \phi_k = (1-\nu)\left(\frac{1}{\gamma^2} - 1\right)\delta_k^2 + (1-\nu)\left(1 - \frac{1}{\gamma}\right)\delta_k \triangleq b_1. \tag{36}$$

We now analyze the case when $\chi(\widehat{x}^k) \geq \zeta\,\delta_k$.
(a) I_k and J_k are both true, i.e., the model f_k is $\kappa = (\kappa_{ef}, \kappa_{ed})$-fully linear and the estimates v_k and $v_{k+1/2}$ are ϵ_F-accurate. With $\eta_1 \in (0,1)$, we have $\chi(\widehat{x}^k) \geq (\kappa_{ed} + \frac{4\kappa_{ef}}{\kappa_{dcp}})\,\delta_{\max}\,\delta_k$. So the condition (16) in Lemma 11 holds, and the trial step s_k lead to a decrease of f such that

$$f(\widehat{x}^k + s^k) - f(\widehat{x}^k) \leq -C_1\,\chi(\widehat{x}^k)\,\delta_k,$$

where $C_1 \triangleq \dfrac{2\kappa_{dcp}\,\kappa_{ef}}{(4\kappa_{ef} + \kappa_{ed}\,\kappa_{dcp})\,\delta_{\max}}$ as defined in Lemma 11. Moreover, since $\epsilon_F \leq \kappa_{ef}$ and

$$\chi_k(\widehat{x}^k) \geq \chi(\widehat{x}^k) - \kappa_{ed}\,\delta_k \geq (\zeta - \kappa_{eg})\,\delta_k \geq \max\left\{\eta_2, \frac{4\,\kappa_{ef}\,\delta_{\max}}{\kappa_{dcp}(1-\eta_1)}\right\}\delta_k,$$

the condition in Lemma 12 holds. Hence k-th iteration is a successful, i.e., $\widehat{x}^{k+1} = \widehat{x}^k + s^k$ and $\delta_{k+1} = \gamma\,\delta_k$. This indicates that

$$\phi_{k+1} - \phi_k \leq -\nu\,C_1\,\chi(\widehat{x}^k)\,\delta_k + (1-\nu)(\gamma^2 - 1)\,\delta_k^2 + (1-\nu)(1-\gamma)\delta_k \triangleq b_2. \tag{37}$$

(b) I_k is true and J_k is false, under which Lemma 11 holds. If the iteration is successful, we can obtain (37). If the iteration is unsuccessful, we can obtain (36). The right-hand side of (37) is strictly smaller than the right-hand side of (36), i.e.,

$$b_2 - b_1 \quad \leq -\nu\,C_1\,\zeta\delta_k^2 + (1-\nu)\delta_k^2\left(\gamma^2 - \frac{1}{\gamma^2}\right) + (1-\nu)\delta_k\left(\frac{1}{\gamma} - \gamma\right) < 0,$$

when $\dfrac{\nu}{1-\nu} \geq \dfrac{4\,\gamma^2}{C_1\zeta}$. Therefore, whether the iteration is successful or not, (36) holds, and we have

$$\phi_{k+1} - \phi_k \leq (1-\nu)\left(\frac{1}{\gamma^2} - 1\right)\delta_k^2 + (1-\nu)\left(1 - \frac{1}{\gamma}\right)\delta_k. \tag{38}$$

(c) I_k is false and J_k is true. If the iteration is successful, since Lemma 13 holds with the condition on ϵ_F, we have

$$f(\widehat{x}^{k+1}) - f(\widehat{x}^k) \leq -C_2\delta_k^2,$$

with $C_2 \triangleq \eta_1\eta_2\frac{\kappa_{dcp}}{\delta_{\max}} - 2\epsilon_F$. Hence, in the case when the iteration is successful, we have

$$\phi_{k+1} - \phi_k \leq -\nu C_2\delta_k^2 + (1-\nu)(\gamma^2 - 1)\,\delta_k^2 + (1-\nu)(1-\gamma)\delta_k \triangleq b_3.$$

If the iteration is unsuccessful, we can obtain (36), which is strictly larger than b_3 as $\dfrac{\nu}{1-\nu} \geq \dfrac{4\,\gamma^2}{C_2}$. Therefore, whether the iteration is successful or not, (36) holds, and we have

$$\phi_{k+1} - \phi_k \leq (1-\nu)\left(\frac{1}{\gamma^2} - 1\right)\delta_k^2 + (1-\nu)\left(1 - \frac{1}{\gamma}\right)\delta_k. \tag{39}$$

(d) I_k and J_k are both false, which can cause the algorithm to accept a bad step and lead to an increase in both f and δ. This part differs from the analysis in [9] due the nonsmoothness of the function h.

$$\begin{aligned} f(\widehat{x}^k + s^k) - f(\widehat{x}^k) = \; & \big(c(\widehat{x}^k + s^k) - c(\widehat{x}^k)\big) + \mathbb{E}_{\widetilde{\varepsilon}}\left[\, g(\widehat{x}^k + s^k, \psi(\widehat{x}^k + s^k) + \widetilde{\varepsilon}) - g(\widehat{x}^k, \psi(\widehat{x}^k) + \widetilde{\varepsilon})\right] \\ & + \mathbb{E}_{\widetilde{\varepsilon}}\left[\max_{j\in\mathcal{J}}\{h_j(\widehat{x}^k + s^k, \psi(\widehat{x}^k + s^k) + \widetilde{\varepsilon})\} - \max_{j\in\mathcal{J}}\{h_j(\widehat{x}^k, \psi(\widehat{x}^k) + \widetilde{\varepsilon})\}\right] \end{aligned}$$

Since both c and g are smooth functions with Lipschitz gradient modulus L_2, by the Taylor's expansion, we have

$$c(\widehat{x}^k + s^k) - c(\widehat{x}^k) \leq \nabla c(\widehat{x}^k + s^k)^\top s_k + \frac{1}{2} L_2\, \delta_k^2 \leq L_1 \delta_k + \frac{L_2}{2}\delta_k^2, \tag{40}$$

and for each realization ε,

$$\begin{aligned} & g(\widehat{x}^k + s^k, \psi(\widehat{x}^k + s^k) + \varepsilon) - g(\widehat{x}^k, \psi(\widehat{x}^k) + \varepsilon) \\ & \leq \nabla_1 g(\widehat{x}^k + s^k, \psi(\widehat{x}^k + s^k) + \varepsilon)^\top s^k + \nabla_2\, g(\widehat{x}^k + s^k, \psi(\widehat{x}^k + s^k) + \varepsilon)^\top (\psi(\widehat{x}^k + s^k) - \psi(\widehat{x}^k)) \\ & + \frac{L_2}{2}\, \Big\| \psi(\widehat{x}^k + s^k) - \psi(\widehat{x}^k) \Big\|^2 + \frac{L_2}{2}\delta_k^2 \\ & \leq L_1(1 + L_1)\, \delta_k + \frac{L_1^2 L_2 + L_2}{2}\, \delta_k^2 \end{aligned} \tag{41}$$

where the last second inequality is derived from the Lipschitz continuity and Lipschitz gradient property of g and ψ. Now we bound the nonsmooth part involving $\max_{j\in\mathcal{J}} h_j(x, \omega)$. For each realization ε, since h_j are convex functions, we have

$$\begin{aligned} & \max_{j\in\mathcal{J}}\{h_j(\widehat{x}^k + s^k, \psi(\widehat{x}^k + s^k) + \varepsilon\,)\} - \max_{j\in\mathcal{J}}\{h_j(\widehat{x}^k, \psi(\widehat{x}^k) + \varepsilon\,)\} \\ & \leq q_1^\top s^k + q_2^\top (\psi(\widehat{x}^k + s^k) - \psi(\widehat{x}^k)) + \frac{L_2}{2}\, \Big\| \psi(\widehat{x}^k + s^k) - \psi(\widehat{x}^k) \Big\|^2 + \frac{L_2}{2}\delta_k^2 \\ & \leq L_1(1 + L_1)\, \delta_k + \frac{L_1^2\, L_2 + L_2}{2}\, \delta_k^2, \end{aligned} \tag{42}$$

where $q_1 \in \partial_1 h(\widehat{x}^k + s^k, \psi(\widehat{x}^k + s^k) + \varepsilon)$, $q_2 \in \partial_2 h(\widehat{x}^k + s^k, \psi(\widehat{x}^k + s^k) + \varepsilon)$. Hence,

$$f(\widehat{x}^k + s^k) - f(\widehat{x}^k) \;\; \leq (3L_1 + 2L_1^2)\delta_k + \frac{2L_1^2\, L_2 + 3L_2}{2}\delta_k^2.$$

If the iteration is successful, with $C_3 \triangleq (3L_1 + 2L_1^2)$, $C_4 \triangleq \dfrac{2L_1^2\, L_2 + 3L_2}{2\zeta}$ and $\chi(\widehat{x}^k) \geq \zeta\, \delta_k$, we have

$$\phi_{k+1} - \phi_k \leq \nu\, C_3\, \delta_k + \nu\, C_4 \chi(\widehat{x}^k)\, \delta_k + (1-\nu)(\gamma^2 - 1)\, \delta_k^2 + (1-\nu)(1-\gamma)\, \delta_k. \tag{43}$$

If the iteration is unsuccessful, we can obtain (36), which is strictly smaller than the right-hand side of (43) as $\dfrac{\nu}{1-\nu} \geq \dfrac{\gamma^2}{C_3}$. Therefore, whether the iteration is successful or not, (43) holds.
Now with the bounds of the differences $\phi_{k+1} - \phi_k$ for all 4 outcomes of $\{I_k, J_k\}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\Phi_{k+1} - \Phi_k \mid \mathcal{F}_{k-1}\right] \\
&\leq \alpha\,\beta\left[-\nu\, C_1\,\chi(\widehat{X}^k)\,\Delta_k + (1-\nu)(\gamma^2-1)\,\Delta_k^2 + (1-\nu)(1-\gamma)\Delta_k\right] \\
&\quad +(\alpha(1-\beta)+\beta(1-\alpha))\left[(1-\nu)\left(\frac{1}{\gamma^2}-1\right)\Delta_k^2 + (1-\nu)\left(1-\frac{1}{\gamma}\right)\Delta_k\right] \\
&\quad +(1-\alpha)(1-\beta)\left[\nu\, C_3\Delta_k + \nu\, C_4\chi(\widehat{X}^k)\,\Delta_k. + (1-\nu)(\gamma^2-1)\Delta_k^2 + (1-\nu)\,(1-\gamma)\,\Delta_k\right] \\
&= (-\nu\,\alpha\beta\, C_1 + (1-\alpha)(1-\beta)\,\nu\, C_4)\,\chi(\widehat{X}^k)\,\Delta_k + (1-\alpha)(1-\beta)\nu\, C_3\,\Delta_k \\
&\quad + \left(\alpha\beta - \frac{1}{\gamma^2}(\alpha(1-\beta)+(1-\alpha)\beta) + (1-\alpha)(1-\beta)\right)(1-\nu)(\gamma^2-1)\Delta_k^2 \\
&\quad - \left(\alpha\beta - \frac{1}{\gamma}\,(\alpha(1-\beta)+(1-\alpha)\beta) + (1-\alpha)(1-\beta)\right)(1-\nu)(\gamma-1)\Delta_k.
\end{aligned}
$$

Since

$$
\alpha\beta - \frac{1}{\gamma^2}(\alpha(1-\beta)+(1-\alpha)\beta) + (1-\alpha)(1-\beta) \leq (\alpha+(1-\alpha))(\beta+(1-\beta)) = 1,
$$

$$
\alpha\beta - \frac{1}{\gamma}\,(\alpha(1-\beta)+(1-\alpha)\beta) + (1-\alpha)(1-\beta) \geq (\alpha-(1-\alpha))(\beta-(1-\beta)) = (2\alpha-1)(2\beta-1),
$$

we have

$$
\begin{aligned}
\mathbb{E}\left[\Phi_{k+1} - \Phi_k \mid \mathcal{F}_{k-1}\right] \;&\leq (-\alpha\beta\, C_1 + (1-\alpha)(1-\beta)\, C_4)\;\nu\,\chi(\widehat{X}^k)\Delta_k + (1-\nu)(\gamma^2-1)\Delta_k^2 \\
&\quad + (\,(1-\alpha)(1-\beta)\nu\, C_3 \;-(2\alpha-1)(2\beta-1)(1-\nu)(\gamma-1))\,\Delta_k.
\end{aligned}
$$

By choosing parameters α and β close to 1 satisfying

$$
\frac{\alpha\beta - \frac{1}{2}}{(1-\alpha)(1-\beta)} \geq \frac{C_4}{C_1}, \text{ and } \quad \frac{(2\alpha-1)(2\beta-1)}{(1-\alpha)(1-\beta)} \geq \frac{\nu\, C_3}{(1-\nu)(\gamma-1)}, \tag{44}
$$

we have

$$
\mathbb{E}\left[\Phi_{k+1} - \Phi_k \mid \mathcal{F}_{k-1}\right] \leq \left(-\frac{1}{2}C_1\nu\zeta + (1-\nu)(\gamma^2-1)\right)\Delta_k^2 \leq -\frac{1}{4}C_1\,\nu\,\zeta\Delta_k^2,
$$

where the last inequality is due to $\dfrac{\nu}{1-\nu} > \dfrac{4\gamma^2}{\zeta C_1}$ from (33). Similar bound can be derived for the case $\chi(\widehat{x}^k) \leq \zeta\delta_k$. By summing it up for all iterations, it can be proved that $\sum_{k=0}^{\infty}\Delta_k^2 < \infty$ almost surely. □

Proof of Theorem 15. As an extension of Theorem 4.16 in [9], we can derive the liminf convergence result that $\liminf_{k\to\infty}\chi(\widehat{X}^k) = 0$ almost surely under the assumptions in Proposition 14.

We omit the proof for this liminf convergence result since it only needs to replace $\|\nabla f(\widehat{X}_k)\|$ in the proof of Theorem 4.16 [9] with $\chi(\widehat{X}_k)$.

Regarding the convergence result that $\lim_{k\to\infty}\chi(\widehat{X}^k)=0$ almost surely, its proof is similar to the proof of Theorem 4.3 in [2], except that we need to make some modifications based on $\chi(\widehat{X}^k)$. Suppose that $\lim_{k\to\infty}\chi(\widehat{X}^k)=0$ does not hold almost surely. Then with a positive probability, there exists $\epsilon>0$ such that $\chi(\widehat{X}^k)>2\epsilon$ holds for infinitely many k. For any $\epsilon>0$, define a sequence of random variables $\{K_\epsilon\}$ consisting of the natural numbers k for which $\chi(\widehat{X}_k)>\epsilon$. Then $\sum_{k\in\{K_\epsilon\}}\Delta_k<\infty$ almost surely by following the proof of Lemma 4.17 in [9] replacing $\|\nabla f(\widehat{X}_k)\|$ with $\chi(\widehat{X}_k)$. We are going to show if such ϵ exists then $\sum_{j\in\{K_\epsilon\}}\Delta_j$ is a divergent sum, and hence, we must have $\lim_{k\to\infty}\chi(\widehat{X}^k)=0$ almost surely.

Since $\liminf_{k\to\infty}\chi(\widehat{X}^k)=0$, there are infinitely many intervals of integers with a positive probability, such that each interval $\{W'+1,\ldots,W''\}$ satisfies: $0<W'<W''$, $\chi(\widehat{X}^{W'})\le\epsilon$, $\chi(\widehat{X}^{W'+1})>\epsilon$, $\chi(\widehat{X}^{W''})>2\epsilon$, and for any integer $w\in(W',W'')$, $\epsilon<\chi(\widehat{X}^w)\le 2\epsilon$. Let $\{(W'_r,W''_r)\}_r$ be an infinite sequence of such intervals. Let (w'_r,w''_r) be the realization of (W'_r,W''_r).

$$\epsilon<\chi(\widehat{x}^{\,w''_r})-\chi(\widehat{x}^{\,w'_r})=\sum_{k=w'_r}^{w''_r-1}\Big(\chi(\widehat{x}^{k+1})-\chi(\widehat{x}^k)\Big)\le\sum_{k=w'_r}^{w''_r-1}\widehat{\chi}_{(L\delta_k)}(\widehat{x}^{k+1})-\chi(\widehat{x}^k),\tag{45}$$

with $L\triangleq 2L_1(L_1+1)$ and $\widehat{\chi}_{(L\delta_k)}(\bullet)$ is defined in (31). From Lemma 17 with the definitions of active index sets $\mathcal{A}(x,\omega)$ for the max function $h(x,\omega)$, for almost every ε, $\mathcal{A}(\widehat{x}^k,\psi(\widehat{x}^k)+\varepsilon)\subseteq\mathcal{A}_{(L\delta_k)}(\widehat{x}^{k+1},\psi(\widehat{x}^{k+1})+\varepsilon)$. Furthermore, let $\hat{d}_k$ be the optimal direction in defining $\widehat{\chi}_{(L\delta_k)}(\widehat{x}^{k+1})$, then we have

$$\begin{aligned}\widehat{\chi}_{(L\delta_k)}(\widehat{x}^{k+1})-\chi(\widehat{x}^k)\;&\le f'(\widehat{x}^k;\hat{d}_k)-\nabla c(\widehat{x}^{k+1})^\top\hat{d}_k\\&\quad-\mathbf{E}_{\widetilde{\varepsilon}}\Big[\,h'_{(L\cdot\delta_k)}\Big((\widehat{x}^{k+1},\psi(\widehat{x}^{k+1})+\widetilde{\varepsilon});(\hat{d}_k,\nabla\psi(\widehat{x}^{k+1})^\top\hat{d}_k)\Big)\Big]\\&\quad-\mathbf{E}_{\widetilde{\varepsilon}}[\,g'\,((\widehat{x}^{k+1},\psi(\widehat{x}^{k+1})+\widetilde{\varepsilon});(\hat{d}_k,\,\nabla\psi(\widehat{x}^{k+1})^\top\,\hat{d}_k))\,]\\&\le 2(L_1\,L_2+L_2)\,\big\|\,\widehat{x}^{k+1}-\widehat{x}^k\,\big\|+\mathbf{E}_{\widetilde{\varepsilon}}\Big[h'((\widehat{x}^k,\psi(\widehat{x}^k)+\widetilde{\varepsilon});(\hat{d}_k,\nabla\psi(\widehat{x}^k)^\top\hat{d}_k))\Big]\\&\quad-\mathbf{E}_{\widetilde{\varepsilon}}\Big[\,h'_{(L\cdot\delta_k)}\Big((\widehat{x}^{k+1},\psi(\widehat{x}^{k+1})+\widetilde{\varepsilon});(\hat{d}_k,\nabla\psi(\widehat{x}^{k+1})^\top\hat{d}_k)\Big)\Big].\end{aligned}$$

Since $\mathcal{A}(\widehat{x}^k,\psi(\widehat{x}^k)+\varepsilon)\subseteq\mathcal{A}_{(L\delta_k)}(\widehat{x}^{k+1},\psi(\widehat{x}^{k+1})+\varepsilon)$, we have

$$\begin{aligned}&\mathbf{E}_{\widetilde{\varepsilon}}\Big[h'((\widehat{x}^k,\psi(\widehat{x}^k)+\widetilde{\varepsilon});(\hat{d}_k,\nabla\psi(\widehat{x}^k)^\top\hat{d}_k))\;-\;h'_{(L\cdot\delta_k)}\Big((\widehat{x}^{k+1},\psi(\widehat{x}^{k+1})+\widetilde{\varepsilon});(\hat{d}_k,\nabla\psi(\widehat{x}^{k+1})^\top\hat{d}_k)\Big)\Big]\\&\le\mathbf{E}_{\widetilde{\varepsilon}}\,\big[\,\max_{j\in\mathcal{J}}\big\|\,\nabla_1 h_j(\widehat{x}^k,\psi(\widehat{x}^k)+\widetilde{\varepsilon})\;-\nabla_1 h_j(\widehat{x}^{k+1},\psi(\widehat{x}^{k+1})+\widetilde{\varepsilon})\,)\,\big\|\,\big]\\&\quad+\mathbf{E}_{\widetilde{\varepsilon}}\,\big[\,\max_{j\in\mathcal{J}}\big\|\,\nabla\psi(\widehat{x}^k)\,\nabla_2 h_j(\widehat{x}^k,\psi(\widehat{x}^k)+\widetilde{\varepsilon})-\nabla\psi(\widehat{x}^{k+1})\,\nabla_2 h_j(\widehat{x}^{k+1},\psi(\widehat{x}^{k+1})+\widetilde{\varepsilon})\,\big\|\,\big]\\&\le L_0\,\big\|\,\widehat{x}^{k+1}-\widehat{x}^k\,\big\|,\end{aligned}$$

for some L_0 depending on L_1 and L_2. This indicates that $\chi_{(L\delta_k)}(\widehat{x}^{k+1})-\chi(\widehat{x}^k)\le L'\,\big\|\,\widehat{x}^{k+1}-\widehat{x}^k\,\big\|$ with $L'=L_0+2(L_1\,L_2+L_2)$. Combining with (45), we derive that for any l,

$$L'\sum_{k=W'_l}^{W''_l-1}\Big\|\,\widehat{x}^{k+1}-\widehat{x}^k\,\Big\|\ge\epsilon,$$

which yields that $\sum_{k\in\{K_\epsilon\}} \Delta_k = \infty$ almost surely, thus contradicts the initial assumption. Hence we have $\lim_{k\to\infty} \chi(\widehat{X}^k) = 0$ almost surely. Furthermore, by Lemma 11.1.2 in [11], this implies that the sequence produced by the CLEO algorithm converges to a directional stationary point almost surely. $\square$