Online Non-parametric Estimation for Nonconvex Stochastic Programming

Shuotao Diao and Suvrajeet Sen [*]

sdiao@usc.edu, s.sen@usc.edu

May, 2022

Abstract

This paper presents a fusion of Stochastic Decomposition and the Majorization-Minimization algorithm (SD-MM) to solve a class of non-convex stochastic programs. The objective function is an expectation of a smooth concave function and a second-stage linear recourse function, which is common in stochastic programming (SP). This extension not only allows new stochastic difference-of-convex (dc) functions but allows new applications in which these two crucial paradigms (SP and dc) can be integrated to provide a more powerful setting for modern applications. This combination also provides an opportunity to study convergence results in a more general setting. Furthermore, with the predictive capability of k nearest neighbors estimation, the proposed algorithm is also extended to solve nonconvex predictive stochastic programming, where the data are present as covariates, and the underlying conditional distribution is unknown. Finally, the computational results of various instances prove the efficiency of the methodology.

1 Introduction

In this paper, we aim to study a fusion of the concepts underlying Stochastic Decomposition with Majorization-Minimization algorithm to solve a class of nonconvex two-stage stochastic programs which can be written as follows:

$$\min_{x \in X} \zeta(x) \triangleq \mathbb{E}_{\tilde{\xi}} \left[F(x, \tilde{\xi}) + H(x, \tilde{\xi}) \right] \tag{1}$$

[*] Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles CA, 90089-0193, USA

where $F : S \times \mathcal{Y} \mapsto \mathbb{R}$ is a L-smooth/differentiable concave function, $S \subset \mathbb{R}^n$ is an open superset of the convex set X, and $H(x, \xi)$ is the minimum cost of the second-stage linear programming problem given the first-stage decision variable, x and, the realization of the random variable, ξ:

$$\begin{aligned} H(x,\xi) = \min_y \; & d^\top y \\ \text{s.t. } & Dy = e(\xi) - C(\xi)x, \\ & y \geq 0, \; y \in \mathbb{R}^{n_2}. \end{aligned} \tag{2}$$

Problem (1) not only includes the classical two-stage stochastic linear program (SLP) as a special case but also two-stage stochastic linear programs with a deterministic nonconvex first-stage objective function. Two-stage stochastic programming has been applied in many areas such as network capacity planning (Sen et al. [42]), water management (Huang and Loucks [21]), logistics (Barbarosoğlu and Arda [3]), power transmission (Phan and Gosh [39]), and others. On the other hand, formulation (1) allows a model in which the first-stage cost function exhibits decreasing returns to scale. For instance, Cafaro and Grossmann [8] use a concave cost function for the economy of scale and model the pipeline installation cost as a concave function of pipeline diameter. Konno and Wijayanayake [24] model the transaction cost as a non-decreasing concave function of the amount of investments.

In addition to the challenge of nonconvexity, we shall also tackle the so-called Predictive Stochastic Programming (PSP) model in which data are available as covariates, (ω, ξ), although the underlying conditional distribution of $\tilde{\xi}$ given $\tilde{\omega} = \omega$ is *unknown*. One example of using covariates information is the hedonic pricing model for housing research (see Owusu-Ansah [37] for an overview). In particular, housing price is often modeled (Kumbhakar and Parmeter [25], Martins-Filho and Bin [35], Witte et al. [49]) as a regression function of dwelling attributes (e.g., land area, number of bedrooms, and so on) and location attributes (e.g., distance central business district, distance to the nearest commercial zone, and so on). It is worth mentioning that researchers often use the terms “predictive”, “contextual”, and “data-driven” to distinguish the SP problems with covariates and unknown conditional distribution from the classic SP problems. Such data are common in non-parametric statistical models, and when combined with decision optimization models, they give rise to a PSP of the following form.

$$\min_{x \in X} \; \zeta_\omega(x) \triangleq \mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}} \left[F(x, \tilde{\xi}) + H(x, \tilde{\xi}) | \tilde{\omega} = \omega \right], \tag{3}$$

where $\mu_{\tilde{\xi}|\omega}$ denotes the conditional distribution of $\tilde{\xi}$ given $\tilde{\omega} = \omega$. We observe that in this case, the objective function is parametrized by ω. To clarify the setup, we note that the optimization is taken only over the decision variable, x, instead of the parameter, ω, we will treat ω as the "subscript" of the objective function. Another reason for making ω as the subscript of the objective function is that ω is fixed throughout the entire optimization process.

Recently, the majorization-minimization (MM) method has attracted considerable attention in the machine learning (ML) and SP communities. One core step of the MM method can be described as follows: a (usually convex) upper-bounding function of the (often nonconvex) original objective is created at the current iterate, and then the upper-bounding function is minimized over the feasible region to obtain the next iterate. Such method has been widely used in difference-of-convex (dc) programs (Le and Pham [26], Pang et al. [38], Tao and An [47]). Mairal [32] proposes a Minimization by Incremental Surrogate Optimization (MISO) method to solve a large sum of continuous functions (possibly with composition and usually related to the loss function or log-likelihood function in machine learning) where he uses a sequence of previously generated surrogate functions to create an upper bound of the objective function. Razaviyayn et al. [41] design a Stochastic Successive Upper-bound Minimization (SSUM) method to solve more general SP models which go beyond finite-sum problems. Similar to the MISO from Mairal [32], SSUM successively uses a sequence of previously generated surrogate functions to construct an upper bound of sample approximation functions. It is worth noting that, in the SSUM algorithm, sample approximation functions are iteratively updated by generating i.i.d. samples, while the upper-bound surrogate functions are updated accordingly. Mairal [31] proposes Stochastic MM (SMM) to solve single-stage nonconvex SP, and Liu et al. [28] extend SMM algorithm for compound SP, in which the isotone outer objective function is nested with the expectations of continuous random functions. More recently, An et al. [1] propose the Stochastic dc algorithm (SDCA) to solve the SP problems where the objective function is dc. All in all, those methods can be regarded as a family of sampled surrogation algorithms. For the readers who are interested in sampled surrogation algorithm, we recommend the monograph by Cui and Pang [10]. As for the two-stage SP with a nonconvex second-stage recouse, Liu et al. [29] propose Regularization-Convexification-Sampling (RCS) approach to solve two-stage SP with bi-parameterized quadratic recourse, and Li and Cui [27] propose a decomposition algorithm which uses an upper approximating function of the partial Moreau envelope of the recourse function for each scenario to construct the surrogate function of the objective function. Note that the RCS approach generates a batch of samples that are independent of the past and only uses the currently generated samples to construct an upper bound of the sample approximation objective function. The novelty of the RCS is that it introduces a quadratic regularizer to make the second-stage objective function strongly convex so that obtaining the dc decomposition of the recourse function by its dual becomes viable.

The Stochastic Decomposition (SD, Higle and Sen [17, 18], and Sen and Liu [43]) algorithm is a successive approximation-decomposition algorithm for solving two-stage stochastic linear programming models. SD successively constructs a local outer (lower-bounding) approximation of the sampled approximation of the second-stage cost-to-go functions by using previously calculated stochastic subgradients at the candidate/incumbent solutions. It is worth noting that SD assumes that the objective function consists of a deterministic linear or convex quadratic first-stage component (see Liu and Sen [30] for the quadratic case) and an expectation of the stochastic second-stage cost-to-go function. The value of the second-stage cost-to-go function at a given feasible first-stage decision is the minimum value of a linear or quadratic programming problem.

One of the main factors motivating this paper is that while the SD algorithm and the MM algorithm share some similarities, they also provide complement strengths within their focal points: SD uses convex lower-bound function (i.e., piecewise linear function) to approximate the sample average approximation of the objective function from below, while the MM algorithm uses a convex upper-bound function to approximate the nonconvex objective function from above. Inspired by these complementary strengths, we aim to design a fusion of SD and MM (SD-MM) to approximate the convex component of the objective function from below and approximate the rest of the nonconvex objective from above in each iteration. It is worth noting that the constructed surrogate function is no longer a global upper bound of the sample average approximation of the objective function.

The use of the feature information to assist decision making has gained increasing attention in the optimization community. Bertsimas and Kallus [6] make a fusion of optimization and machine learning to propose a new type of SAA-based models with predictive power and then apply the model in the inventory management problem. In particular, they propose a non-parametric approximation of (3) as follows,

$$\min_{x \in X} \sum_{i=1}^{N} v_{N,i}(\omega) \left[F(x, \tilde{\xi}) + H(x, \tilde{\xi}) \right], \tag{4}$$

where $v_{N,i}(\omega)$ is calculated by some non-parametric estimation method (e.g., in the case of kNN estimation, $v_{N,i}(\omega) = \frac{1}{k_N} \mathbb{I}(\omega_i \in \mathcal{S}(k_N, \omega; \{\omega_i\}_i^N))$ and $\mathcal{S}(k_N, \omega; \{\omega_i\}_i^N)$ denotes the kNN set of ω from the set $\{\omega_i\}_i^N$). Additionally, Bertsimas and McCord [7] further extends the framework to multistage SP. Ban and Rudin [2] propose a feature-based newsvendor problem. Elmachtoub and Grigas [14] propose a Smart "Predict, then Optimize" framework to solve contextual stochastic optimization. Kanna et al. [23] analyze various versions of empirical residuals-based sample average approximation under assumption of the homoscedasticity and then analyze the heteroscedastic case in [22]. Hu et al. [20] study noise dependent/independent rate of regret bound of the estimate-and-then-optimization model and induced empirical-risk-minimization (i.e., simultaneous estimation and optimization) model for solving contextual linear optimization. In all, they generally focus on the SAA-based modelling approaches, which is generally "static". On the other hand, Diao and Sen [12] design an online first-order method by simultaneously refining subgradient estimation and estimated solutions, which they refer to as Learning Enabled Optimization with Non-parametric Approximation (LEON). In this paper, we focus on extending LEON to nonconvex SP. Indeed, we shall design an online SP algorithms with "predictive" power and study an efficient way to reuse the previous non-parametric estimations for enhancing future estimations in a more "dynamic" way. Therefore, we refer to this type of methodology as PSP in this paper.

In this paper, we shall answer the following questions:(i) How should one ensure that locally lower-bounding function of the second-stage cost-to-go function is "good" enough per iteration so that the algorithm converges? (ii) Does any accumulation point of the estimated solutions generated by SD-MM converge to a d-stationary point of problem (1)? (iii) How to synthesize the SD-MM algorithm with kNN estimation (NSD-MM) to solve two-stage nonconvex PSP problems in (3) and how to efficiently reuse previous kNN estimations for future estimations by simply doing some re-scaling?

As suggested in the preceding paragraph, the contributions of this paper are three-fold: (1) We propose a mix of SD and MM to solve nonconvex SP. (2) We extend the boundary of SD and provide a more general methodology and proof by identifying an associated supermartingale of the iterates. It is worth noting that such techniques are relatively uncommon in the SP literature, and are likely to open up a new approach for designing algorithms by exploiting the potential we present. (3) We propose a stochastic MM method based on non-parametric estimation for solving a nonconvex predictive SP and also equip SD-MM with kNN estimation to solve nonconvex predictive SP problems.

The paper is organized as follows. In section 2, we propose a design of a fusion of the SD algorithm and the MM algorithm to solve unparametrized version of problem in (1). In section 3, we design a kNN extension of SD-MM algorithm to solve the PSP problem in (3). Finally, we apply our proposed algorithm and its non-parametric extension to numerically solve a class of two-stage stochastic program and two-stage predictive stochastic program in section 4.

1.1 Technical Preliminaries and Notations

Without further specification, we shall let $\|\cdot\|$ denote the Euclidean norm of the vector and spectral norm of the matrix. Let $\tilde{\xi} : \Omega \mapsto \mathcal{Y} \subset \mathbb{R}^{m_2}$ be a random vector built upon the probability space $(\Omega, \Sigma_\Omega, \mathbb{P})$. We let $\mu_{\tilde{\xi}}$ denote the distribution of the random vector $\tilde{\xi}$ and let ξ denote a realization of $\tilde{\xi}$. When we say "for almost every $\xi \in \mathcal{Y}$", it means for $\mu_{\tilde{\xi}}$ almost every $\xi \in \mathcal{Y}$. Throughout, we assume that X is a compact convex set.

Definition 1. *A function $F : \mathbb{R}^n \mapsto \mathbb{R}$ is a L-smooth function if it is differentiable and its gradient is L-Lipschitz continuous on $\mathbb{R}^n$ (i.e., $\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\|$, for all $x, y \in \mathbb{R}^n$).*

We define a class of surrogate function of $F(x, \xi)$ near x' below.

Definition 2. *Let $F : S \times \mathcal{Y} \mapsto \mathbb{R}$ be a L-smooth/differentiable concave function on X, where S is a superset of X. $G(x, \xi; x')$ is a convex surrogate function of $F(x, \xi)$ near $x' \in X$ which satisfies the following conditions:*

M1 $G(x', \xi; x') = F(x', \xi)$.

M2 $G(x, \xi; x') \geq F(x, \xi),\ \forall\ x \in X$.

M3 $G(\lambda x^1 + (1-\lambda)x^2, \xi; x') \leq \lambda G(x^1, \xi; x') + (1-\lambda)G(x^2, \xi; x'),\ \forall\ x^1, x^2 \in X,\ \forall\ \lambda \in [0, 1]$.

Definition 3. *Let $\zeta : S \mapsto \mathbb{R}$, where S is a convex set of $\mathbb{R}^n$. Suppose that the directional derivative of f at $x \in X$ in the direction of $d \in \mathbb{R}^n$ exists and it is defined below:*

$$\zeta'(x; d) = \lim_{\tau \downarrow 0} \frac{\zeta(x + \tau d) - \zeta(x)}{\tau}.$$

Definition 4. *$\bar{x} \in X$ is a d(irectional)-stationary point of (1) if*

$$\zeta'(\bar{x}; x - \bar{x}) \geq 0,\ \forall\ x \in X.$$

In the PSP, Let $\mathcal{X}$ be a Borel subset of $\mathbb{R}^{m_1}$ and let $\mathcal{Y}$ be a Borel subset of $\mathbb{R}^{m_2}$. Let $(\Omega, \Sigma_\Omega, \mathbb{P})$ be the probability space of the correlated continuous random vectors $\tilde{\omega}$ and $\tilde{\xi}$ taking the values in the measurable spaces, $(\mathcal{X}, \Sigma_{\mathcal{X}})$ and $(\mathcal{Y}, \Sigma_{\mathcal{Y}})$, respectively. In particular, the tuple $(\tilde{\omega}, \tilde{\xi}) : \Omega \mapsto \mathcal{X} \times \mathcal{Y}$ is a random vector taking values in the product space $(\mathcal{X} \times \mathcal{Y}, \Sigma_{\mathcal{X}} \otimes \Sigma_{\mathcal{Y}})$. Let the joint distribution of $(\tilde{\omega}, \tilde{\xi})$ be $\mu_{\tilde{\omega}, \tilde{\xi}}$. Let $\mu_{\tilde{\omega}}$ and $\mu_{\tilde{\xi}}$ denote the marginal distribution functions of $\tilde{\omega}$ and $\tilde{\xi}$, respectively. Let $\mu_{\tilde{\xi}|\omega}$ denote the conditional distribution of $\tilde{\xi}$ given $\tilde{\omega} = \omega$. We refer to Çınlar [9] and Durrett [13] for readers who are interested in joint distributions and conditional distributions. As a result, for the problem parametrized by ω, we shall define the directional derivative of the first argument, x, as follows.

Definition 5. *Let $\zeta_\omega(x) : S \times S_\omega \mapsto \mathbb{R}$, where S is a convex set of $\mathbb{R}^n$ and S_ω is a subset of $\mathbb{R}^{n_\omega}$. For a given $\omega \in S_\omega$, suppose that the directional derivative of $\zeta_\omega(\cdot)$ at $x \in X$ in the direction of $d \in \mathbb{R}^n$ exists and it is defined below:*

$$\zeta'_\omega(x; d) = \lim_{\tau \downarrow 0} \frac{\zeta_\omega(x + \tau d) - \zeta_\omega(x)}{\tau}.$$

Definition 6. *$\bar{x}$ is a d(irectional)-stationary point of (3) if*

$$\zeta'_\omega(\bar{x}; x - \bar{x}) \geq 0, \ \forall \ x \in X,$$

2 SD-MM for SP

In this section, we consider the following (nonconvex) SP problem:

$$\min_{x \in X} \ \mathbb{E}_{\tilde{\xi}} \left[F(x, \tilde{\xi}) + H(x, \tilde{\xi}) \right] \tag{5}$$

where $F : S \times \mathcal{Y} \mapsto \mathbb{R}$ is a L-smooth/differentiable concave function on X, where S is a superset of X, and $H(x, \xi)$ is the minimum cost of the second-stage linear programming problem as defined in (2). We shall make the following assumptions:

A1 X is a convex compact (possibly polyhedron) set.

A2 For almost every $\xi \in \mathcal{Y}$, $F(\cdot, \xi)$ is differentiable on S and there exists a finite $\kappa_f > 0$ such that $\|\nabla F(x, \xi)\| \leq \kappa_f$ for all $x \in X$.

A3 There exists $\bar{f} \in (0, \infty)$ such that $|F(x, \xi)| \leq \bar{f}$ on X for almost every $\xi \in \mathcal{Y}$.

A4 For almost every $\xi \in \mathcal{Y}$, the set $\{y : Dy = e(\xi) - C(\xi)x, y \geq 0\} \neq \emptyset$ for all $x \in X$. Furthermore, $H(x, \xi) \geq 0$ for all $x \in X$ and almost every $\xi \in \mathcal{Y}$.

A5 The subdifferential of H with respect to x is nonempty on X for almost every $\xi \in \mathcal{Y}$ (i.e., $\partial_x H(x, \xi) \neq \emptyset$).

A6 There exists $\kappa_e, \kappa_C \in (0, \infty)$ such that $\|e(\xi)\| \leq \kappa_e$ and $\|C(\xi)\| \leq \kappa_C$ for almost all $\xi \in \mathcal{Y}$. The dual feasible region of the second-stage problem, $\{\pi : \pi^\top D \leq d\}$ is bounded.

Assumptions A1 - A3 are necessary when we apply the convergence rate of the sample average approximation of the objective function. Assumptions A4 - A6 are common in two-stage stochastic linear program literature, in which Assumption A4 corresponds to the relative complete recourse assumption, Assumptions A5 and A6 altogether ensures that the subdifferential of the recourse function exists and is bounded for all $x \in X$. For the reminder of this section, we introduce the following notations:

ℓ : iteration number of the outer loop, ν : iteration number of the inner loop

ν_ℓ : number of inner loops in the outer iteration ℓ

$\{x^\ell\}$: sequence of incumbents generated by the outer loop

$\{x_\nu^{\frac{1}{2},\ell}\}$: sequence of candidates generated by the inner loops

$$f_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell} F(x,\xi_i),\ h_\ell(x) = \frac{1}{\ell}\sum_{i=1}^{\ell} H(x,\xi_i),\ \zeta_\ell(x) = f_\ell(x) + h_\ell(x)$$

$$f(x) = \mathbb{E}_{\tilde{\xi}}[F(x,\tilde{\xi})],\ h(x) = \mathbb{E}_{\tilde{\xi}}[H(x,\tilde{\xi})]$$

$$\zeta(x) = f(x) + h(x),\ \|\zeta - \zeta_\ell\|_\infty = \sup_{x \in X} |\zeta(x) - \zeta_\ell(x)|$$

$\hat{h}_{\ell,\nu}(x)$: piecewise linear approximation of $h_\ell(x)$ in the ν^{th} inner iteration of the ℓ^{th} outer loop

$$g_\ell(x;x') = \frac{1}{\ell}\sum_{i=1}^{\ell} G(x,\xi_i;x')$$

We further define that $f_0 \equiv 0$, $h_0 \equiv 0$, $\zeta_0 \equiv 0$.

2.1 Algorithm Design

The SD-MM algorithm consists of inner loops for successively refining the lower-bound approximation of the sample average of $\mathbb{E}_{\tilde{\xi}}[H(x,\tilde{\xi})]$ and outer loops for finding a sequence of incumbent solutions. Note that the dual of the linear program in (2) is $\max\{\pi^\top(e(\xi) - C(\xi)x) : \pi^\top D \leq d\}$. We shall iteratively compute the dual extreme points of the sampled second-stage problem to construct the piecewise linear approximation of $\mathbb{E}_{\tilde{\xi}}[H(x,\tilde{\xi})]$. The design of inner of loop is inspired by Higle and Sen [17, 18] and Philpott and Guan [40] since the sample average of the $\mathbb{E}_{\tilde{\xi}}[H(x,\tilde{\xi})]$ in iteration ℓ can be regarded as a finite sum of piecewise linear function of x each with finitely many pieces. Therefore, this allows us to iteratively sample the active pieces of the sample approximation of $\mathbb{E}_{\tilde{\xi}}[H(x,\tilde{\xi})]$ and include it in the lower-bound approximation function of $\mathbb{E}_{\tilde{\xi}}[H(x,\tilde{\xi})]$. The stopping criterion of the inner loop ensures that the difference between the lower bounding approximation function and the sample average of the second-stage recourse function in the next incumbent solution is bounded by a certain portion of the proximal term. Therefore, it ensures a relative descent (based on the value of the sample approximation of the objective function).

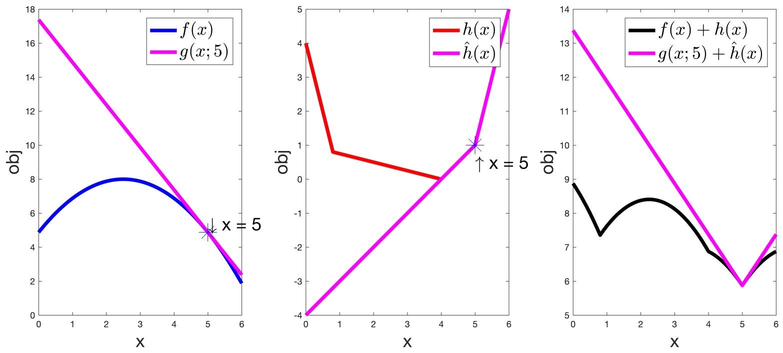

Figure 1: Illustration of the SD-MM methodology. The first figure (starting from the left hand side) illustrates that SD-MM approximates the concave function $f(x)$ from above. The second figure illustrates that SD-MM approximates the convex function $h(x)$ from below. The last figure shows that the sum of the two function approximations becomes a local approximation of $f(x) + h(x)$.

Remark 1: In the candidate update step, the optimization problem can be reformulated as

$$\begin{aligned} \min_x \; & g_\ell(x;x^\ell) + \eta + \frac{c}{2}\|x - x^\ell\|^2 \\ \text{s.t. } & \alpha_t^\ell + (\beta_t^\ell)^\top x \le \eta, \; t \in \mathcal{T}_{\ell,\nu} \\ & x \in X \end{aligned} \tag{6}$$

Remark 2: Since $\hat{h}_\ell$ is the outer approximation of h_ℓ, we have $\hat{h}_\ell(x) \le h_\ell(x)$ for all $x \in X$, which will be formally shown in the next subsection. Since h_ℓ is a piecewise linear function of x with finitely many pieces, the number of iterations in the inner loop is finite for each ℓ. In the worst case, the inner loop in the ℓ^{th} outer loop will recover all pieces of h_ℓ. **Piecewise Linear Approximation Update I** adds a minorant of h_ℓ at x^ℓ and update the previous minorants so that the updated approximation function is an outer approximation of h_ℓ.

Remark 3: The minorant pruning is necessary since it limits the number of pieces in the approximation of $h(x)$ at the beginning of each outer iteration. As a result, it ensures the efficiency of the local approximation of $h(x)$.

Algorithm 1 SD-MM

```
(Initialization) Pick $x^1 \in X$, $c > 0$, and $L > 0$ (number of iterations). Let $\nu = 0$, and $\nu_0 = 0$. Set $\hat{h}_{0,1}(x) \equiv 1$.
for $\ell = 1, 2, \dots, L$ do                                                    Outer Loop
    Generate $\xi_\ell \sim \mu_{\tilde{\xi}}$, which is independent from the past samples.
    Compute $\pi_i^\ell \in \arg\max\{\pi^\top(e(\xi_i) - C(\xi_i)x^\ell) : \pi^\top D \leq d\}$ for $i = 1, 2, \dots, \ell$.
    Compute $u_\ell(x^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell} C(\xi_i)^\top \pi_i^\ell$
    if $\ell > 1$ then
        Perform Minorant Pruning I/II for $\hat{h}_{\ell-1,\nu_{\ell-1}+1}(x)$
    end if
    Set $\underbrace{\hat{h}_{\ell,1}(x) \leftarrow \max\{\frac{\ell-1}{\ell}\hat{h}_{\ell-1,\nu_{\ell-1}+1}(x), h_\ell(x^\ell) + u_\ell(x^\ell)^\top(x - x^\ell)\}}_{\text{Piecewise Linear Approximation Update I}}$.
    Set $\nu \leftarrow 0$.
    do                                                                            Inner Loop
        Set $\nu \leftarrow \nu + 1$.
        Candidate Update: $x_\nu^{\frac{1}{2},\ell+1} = \arg\min_{x \in X} g_\ell(x; x^\ell) + \hat{h}_{\ell,\nu}(x) + \frac{c}{2}\|x - x^\ell\|^2$.
        Compute $\pi_{\nu,i}^{\frac{1}{2},\ell+1} \in \arg\max\{\pi^\top(e(\xi_i) - C(\xi_i)x_\nu^{\frac{1}{2},\ell+1}) : \pi^\top D \leq d\}$ for $i = 1, 2, \dots, \ell$
        Compute $u_\ell(x_\nu^{\frac{1}{2},\ell+1}) = \frac{1}{\ell}\sum_{i=1}^{\ell} C(\xi_i)^\top \pi_{\nu,i}^{\frac{1}{2},\ell+1}$.
        Set $\underbrace{\hat{h}_{\ell,\nu+1}(x) \leftarrow \max\{\hat{h}_{\ell,\nu}(x), h_\ell(x_\nu^{\frac{1}{2},\ell+1}) + u_\ell(x_\nu^{\frac{1}{2},\ell+1})^\top(x - x_\nu^{\frac{1}{2},\ell+1})\}}_{\text{Piecewise Linear Approximation Update II}}$.
    while $h_\ell(x_\nu^{\frac{1}{2},\ell+1}) - \hat{h}_{\ell,\nu}(x_\nu^{\frac{1}{2},\ell+1}) > \frac{c}{4}\|x_\nu^{\frac{1}{2},\ell+1} - x^\ell\|^2$          End Inner Loop
    Set $x^{\ell+1} \leftarrow x_\nu^{\frac{1}{2},\ell+1}$ and $\nu_\ell \leftarrow \nu$.
end for                                                                           End Outer Loop
```

Algorithm 2 Minorant Pruning I

```
Input: $\hat{h}_{\ell-1,\nu_{\ell-1}+1}(x)$
if number of minorants in $\hat{h}_{\ell-1,\nu+1}(x)$ is greater than $N$ then
    Remove the earliest minorants so that the number of minorants in $\hat{h}_{\ell-1,\nu+1}(x)$ equal to $N$
end if
Output: $\hat{h}_{\ell-1,\nu_{\ell-1}+1}(x)$
```

Algorithm 3 Minorant Pruning II

```
Input: $x^\ell$, $h_{\ell-1}(x^\ell)$, $h_{\ell-1}(x^{\ell-1})$, $u_{\ell-1}(x^\ell)$, $u_{\ell-1}(x^{\ell-1})$, $\hat{h}_{\ell-1,\nu_{\ell-1}}(x)$
Remove the minorants from $\hat{h}_{\ell-1,\nu_{\ell-1}}(x)$ that are not active at $x_{\nu_{\ell-1}}^{\frac{1}{2},\ell}$ (i.e., This can be achieved by solving the Candidate Update in its alternative formulation in (6) and recording the Lagrange multipliers of minorants of $\hat{h}_{\ell-1,\nu_{\ell-1}+1}(x)$ after processing the Candidate Update and then deleting the minornats whose Lagrange multipliers are 0).
Set
    $\hat{h}_{\ell-1,\nu_{\ell-1}+1}(x) = \max\{\hat{h}_{\ell-1,\nu_{\ell-1}}(x), \underbrace{h_{\ell-1}(x^\ell) + u_{\ell-1}(x^\ell)^\top(x - x^\ell)}_{\text{Minorant at } x^\ell}$
        $, \underbrace{h_{\ell-1}(x^{\ell-1}) + u_{\ell-1}(x^{\ell-1})^\top(x - x^{\ell-1})}_{\text{Minorant at } x^{\ell-1}}\}$
Output: $\hat{h}_{\ell-1,\nu_{\ell-1}+1}(x)$
```

2.2 Convergence Analysis

First, we show that $\hat{h}_\ell$ is an outer approximation of h_ℓ on X (i.e., $\hat{h}_\ell(x) \leq h_\ell(x)$ for all $x \in X$) for all $\ell \geq 1$. It is worth noting that the similar result (without inner-loop updates) is first shown by Higle and Sen [17]. Second, we show that $\{\|x^{\ell+1} - x^\ell\|\}_\ell$ converges to 0 with probability one. Finally, we show that any accumulation of $\{x^\ell\}_\ell$ generated by Algorithm 1 is a d-stationary point of (1).

Proposition 1. *Suppose that Assumptions A1 - A6 hold. With probability one, $\hat{h}_{\ell,\nu}(x) \leq h_\ell(x)$ for all $x \in X$ and for all $\ell \geq 1$ and $\nu \geq 1$.*

Proof. By the lower bound of $H(x, \xi)$ in Assumption A4 and the direct use of subgradients to construct the minorants, the results follow. ■

It is worth noting that $g_\ell(x; x^\ell) + \hat{h}_{\ell,\nu}(x)$ constructed by the SD-MM is neither an upper bound nor lower bound of the function $\zeta_\ell(x)$, which is vastly different from the most of the MM literature. With the design of the inner loop, the descent in terms of the function values of ζ_ℓ at x^ℓ and $x^{\ell+1}$ is ensured, which is formally stated in the following proposition.

Proposition 2. *Suppose Assumptions A1 - A6 hold. Let $\{x^\ell\}_\ell$ be the sequence generated by Algorithm 1. Then the following holds:*

$$f_\ell(x^{\ell+1}) + h_\ell(x^{\ell+1}) + \frac{c}{4}\|x^{\ell+1} - x^\ell\|^2 \leq f_\ell(x^\ell) + h_\ell(x^\ell). \tag{7}$$

Proof. By the **candidate update** of Algorithm 1, we have

$$g_\ell(x^{\ell+1}; x^\ell) + \hat{h}_{\ell,\nu_\ell}(x^{\ell+1}) + \frac{c}{2}\|x^{\ell+1} - x^\ell\|^2 \leq g_\ell(x^\ell; x^\ell) + \hat{h}_{\ell,\nu_\ell}(x^\ell) = f_\ell(x^\ell) + h_\ell(x^\ell) \tag{8}$$

The equality in (8) holds because of the property of the majorant function (M1) and **Piecewise Linear Approximation I** in Step 2 of Algorithm 1.

On the other hand, by the property of the majorant function (M2), we have

$$g_\ell(x^{\ell+1}; x^\ell) \geq f_\ell(x^{\ell+1}). \tag{9}$$

According to Algorithm 1, after the inner loop is terminated, we have

$$\hat{h}_{\ell,\nu_\ell}(x^{\ell+1}) \geq h_\ell(x^{\ell+1}) - \frac{c}{4}\|x^{\ell+1} - x^\ell\|^2 \tag{10}$$

Thus, the final result follows from the combination of (8), (9), and (10). ■

The next proposition establishes the key convergence of the SD-MM algorithm.

Proposition 3. *Suppose Assumptions A1 - A6 hold, then with probability one, $\{f_{\ell-1}(x^\ell) + h_{\ell-1}(x^\ell)\}_\ell$ converges and $\sum_{\ell=1}^{\infty} \|x^{\ell+1} - x^\ell\|^2 < \infty$.*

Proof. Note that x^1 is given in the initialization of the algorithm, $f_0(x^1) + h_0(x^1) = 0$, and $\|x^2 - x^1\|^2 < \infty$ *w.p.*1, by the definition of f_0, h_0, and boundedness of X. So it is equivalent to show $\{f_{\ell-1}(x^\ell) + h_{\ell-1}(x^\ell)\}_{\ell \geq 2}$ and $\sum_{\ell=2}^{\infty} \|x^{\ell+1} - x^\ell\|^2 < \infty$ *w.p.*1. To the end of this proof, we consider that $\ell \geq 2$.

$$
\begin{aligned}
& f_\ell(x^{\ell+1}) + h_\ell(x^{\ell+1}) - \big[f_{\ell-1}(x^\ell) + h_{\ell-1}(x^\ell)\big] && \text{(11a)}\\
& = f_\ell(x^{\ell+1}) + h_\ell(x^{\ell+1}) - \big[f_\ell(x^\ell) + h_\ell(x^\ell)\big] \\
& \ + \big[f_\ell(x^\ell) + h_\ell(x^\ell)\big] - \big[f_{\ell-1}(x^\ell) + h_{\ell-1}(x^\ell)\big] \\
& \leq -\frac{c}{4}\|x^{\ell+1} - x^\ell\|^2 + \big[f_\ell(x^\ell) + h_\ell(x^\ell)\big] - \big[f_{\ell-1}(x^\ell) + h_{\ell-1}(x^\ell)\big] && \text{(11b)}\\
& = -\frac{c}{4}\|x^{\ell+1} - x^\ell\|^2 + \frac{-1}{\ell(\ell-1)}\sum_{i=1}^{\ell-1}[F(x^\ell,\xi_i) + H(x^\ell,\xi_i)] \\
& \ + \frac{1}{\ell}[F(x^\ell,\xi_\ell) + H(x^\ell,\xi_\ell)] \\
& = -\frac{c}{4}\|x^{\ell+1} - x^\ell\|^2 + \frac{-1}{\ell}\zeta_{\ell-1}(x^\ell) + \frac{1}{\ell}[F(x^\ell,\xi_\ell) + H(x^\ell,\xi_\ell)]. && \text{(11c)}
\end{aligned}
$$

Inequality in (11b) holds by Proposition 1. Let $\mathcal{F}_\ell = \sigma(x^1, x^2, \ldots, x^\ell, \xi_1, \xi_2, \ldots, \xi_{\ell-1})$ denote the natural history (filtration) before the outer iteration ℓ. By taking the conditional expectation of (11a) and (11c) with respect to $\mathcal{F}_\ell$, we have

$$
\begin{aligned}
& \mathbb{E}[f_\ell(x^{\ell+1}) + h_\ell(x^{\ell+1})|\mathcal{F}_\ell] \\
& \leq f_{\ell-1}(x^\ell) + h_{\ell-1}(x^\ell) - \frac{c}{4}\mathbb{E}[\|x^{\ell+1} - x^\ell\|^2 \mid \mathcal{F}_\ell] + \frac{-1}{\ell}\zeta_{\ell-1}(x^\ell) + \frac{1}{\ell}\zeta(x^\ell) && \text{(12)}\\
& \leq f_{\ell-1}(x^\ell) + h_{\ell-1}(x^\ell) - \frac{c}{4}\mathbb{E}[\|x^{\ell+1} - x^\ell\|^2 \mid \mathcal{F}_\ell] + \frac{\|\zeta - \zeta_{\ell-1}\|_\infty}{\ell}
\end{aligned}
$$

The first inequality of (12) holds because $\mathbb{E}[F(x^\ell,\xi_\ell)|\mathcal{F}_\ell] = \mathbb{E}_{\tilde{\xi}}[F(x^\ell,\tilde{\xi})] = f(x^\ell)$ and $\mathbb{E}[H(x^\ell,\xi_\ell)|\mathcal{F}_\ell] = \mathbb{E}_{\tilde{\xi}}[H(x^\ell,\tilde{\xi})] = h(x^\ell)$. The second inequality of (12) holds because $\frac{\zeta(x^\ell) - \zeta_{\ell-1}(x^\ell)}{\ell} \leq \frac{\|\zeta - \zeta_{\ell-1}\|_\infty}{\ell}$.
By Donsker Theorem (see Lemma 7 in [33]), there exists $K \in (0,\infty)$ such that

$$
\mathbb{E}[\|\zeta - \zeta_{\ell-1}\|_\infty] \leq \frac{K}{(\ell-1)^{\frac{1}{2}}}, \text{for } \ell \geq 2
$$

Hence,

$$\sum_{\ell=2}^{\infty}\mathbb{E}\left[\frac{\|\zeta-\zeta_{\ell-1}\|_{\infty}}{\ell}\right]\leq\sum_{\ell=2}^{\infty}\frac{K}{(\ell-1)^{\frac{3}{2}}}<\infty \tag{13}$$

which implies that $\sum_{\ell=2}^{\infty}\frac{\|\zeta-\zeta_{\ell-1}\|_{\infty}}{\ell}<\infty$ $w.p.1$. We substract by $\bar{f}$ on both sides of (12) , then $f_{\ell-1}(x^{\ell})+h_{\ell-1}(x^{\ell})-\bar{f}\geq 0$ for all $\ell\geq 2$. Hence, by Supermartingale Convergence Theorem (Proposition 2 in Bertsekas [5]), since $\sum_{\ell=2}^{\infty}\frac{\|\zeta-\zeta_{\ell-1}\|_{\infty}}{\ell}<\infty$ $w.p.1$, we have $\{f_{\ell-1}(x^{\ell})+h_{\ell-1}(x^{\ell})-\bar{f}\}_{\ell\geq 2}$ converges and hence $\{f_{\ell-1}(x^{\ell})+h_{\ell-1}(x^{\ell})\}_{\ell\geq 2}$ converges with probability one. Furthermore, Supermartingale Convergence Theorem implies that $\sum_{\ell=2}^{\infty}\mathbb{E}[\|x^{\ell+1}-x^{\ell}\|^2\mid\mathcal{F}_{\ell}]<\infty$ $w.p.1$ which further implies that $\sum_{\ell=2}^{\infty}\mathbb{E}[\|x^{\ell+1}-x^{\ell}\|^2]<\infty$ and hence $\sum_{\ell=2}^{\infty}\|x^{\ell+1}-x^{\ell}\|^2<\infty$ $w.p.1$. ■

Remark 1: Identifying the "descent" relation in (11) of Proposition 3 is inspired by proof of Lemma 1 in [41]. It is worth noting that the analysis of almost all the sampled surrogation algorithms with incremental sampling involves using the supermartingale convergence theorem (or its variation) and the rate of convergence of the sampled objective function. The novel part of the SD-MM algorithm is that it designs a double-loop structure to incorporate the decomposition-based methods into this unifying framework.

Remark 2: The proof of Proposition 3 is straightforward to be extended to prove the convergence of the mini-batch version of the SD-MM algorithm. For the rest of this section, we stick with the base version of the SD-MM algorithm (i.e., the batch size is 1), but it is straightforward to extend the proof to the mini-batch version of the SD-MM algorithm.

The proof of the following proposition is inspired by the proof of Lemma 4 in Higle and Sen [18] but the conclusion is stronger (i.e., the argument on the entire sequence versus the argument on a subsequence).

Proposition 4. *Suppose Assumptions A1 - A6 hold. Then with probability one, any accumulation point of* $\{x^{\ell}\}_{\ell}$ *(i.e.,* $\lim_{\ell(\in\mathcal{L})\to\infty}x^{\ell}=x^{\infty}$*) satisfies the following relations:*

$$\lim_{\ell(\in\mathcal{L})\to\infty}x^{\ell+1}=\lim_{\ell(\in\mathcal{L})\to\infty}x^{\ell}=x^{\infty}\in X, \tag{14}$$

and

$$\limsup_{\ell(\in\mathcal{L})\to\infty}\hat{h}'_{\ell,\nu_{\ell}}(x^{\ell+1};x-x^{\ell+1})\leq h'(x^{\infty};x-x^{\infty}),\ \forall\ x\in X. \tag{15}$$

Proof. Since X is compact and $\{x^{\ell}\}_{\ell}\subset X$, the accumulation point of $\{x^{\ell}\}_{\ell}$ exists, say $\lim_{\ell(\in\mathcal{L})\to\infty}x^{\ell}=x^{\infty}\in X$. By Proposition 3, we have $\sum_{\ell=1}^{\infty}\|x^{\ell+1}-x^{\ell}\|<\infty$ $w.p.1$, which implies that $\lim_{\ell\to\infty}\|x^{\ell+1}-x^{\ell}\|=0$ and thus $\lim_{\ell(\in\mathcal{L})\to\infty}x^{\ell+1}=\lim_{\ell(\in\mathcal{L})\to\infty}x^{\ell}=x^{\infty}$.

Pick $x \in X$. Let $\{(\alpha_t^\ell, \beta_t^\ell)\}_{t \in \mathcal{T}_{\ell,\nu_\ell}}$ denote the collection of minorant coefficients of $\hat{h}_{\ell,\nu_\ell}(x)$ at the end of the ℓ^{th} outer loop. That is, $\hat{h}_{\ell,\nu_\ell}(x) = \max\limits_{t \in \mathcal{T}_{\ell,\nu_\ell}} \{\alpha_t^\ell + (\beta_t^\ell)^\top x\}$. Let

$$(\alpha_{t_\ell}^\ell, \beta_{t_\ell}^\ell) \in \arg\max\{(\beta_t^\ell)^\top (x - x^{\ell+1}) \mid \alpha_t^\ell + (\beta_t^\ell)^\top x^{\ell+1} = \hat{h}_{\ell,\nu_\ell}(x^{\ell+1}),\ t \in \mathcal{T}_{\ell,\nu_\ell}\}.$$

By Danskin's theorem, $\hat{h}'_{\ell,\nu_\ell}(x^{\ell+1}; x - x^{\ell+1}) = (\beta_{t_\ell}^\ell)^\top (x - x^{\ell+1})$. By Assumption A4, $\{\beta_{t_\ell}^\ell\}_{\ell \in \mathcal{L}}$ is bounded. Hence, $\limsup\limits_{\ell(\in\mathcal{L})\to\infty} \hat{h}'_{\ell,\nu_\ell}(x^{\ell+1}; x - x^{\ell+1}) < \infty$ and there exists a further subsequence of $\mathcal{L}$, say $\mathcal{L}_0$, such that

$$\begin{cases} \lim\limits_{\ell(\in\mathcal{L}_0)\to\infty} (\alpha_{t_\ell}^\ell, \beta_{t_\ell}^\ell) = (\alpha^\infty, \beta^\infty), \\ \limsup\limits_{\ell(\in\mathcal{L})\to\infty} \hat{h}'_{\ell,\nu_\ell}(x^{\ell+1}; x - x^{\ell+1}) = \lim\limits_{\ell(\in\mathcal{L}_0)\to\infty} (\beta_{t_\ell}^\ell)^\top (x - x^{\ell+1}) = (\beta^\infty)^\top (x - x^\infty). \end{cases}$$

By Proposition 1 and the terminating criterion of the inner loop, we have

$$h_\ell(x^{\ell+1}) - \frac{c}{4}\|x^{\ell+1} - x^\ell\|^2 \leq \hat{h}_{\ell,\nu_\ell}(x^{\ell+1}) = \alpha_{t_\ell}^\ell + (\beta_{t_\ell}^\ell)^\top x^{\ell+1} \leq h_\ell(x^{\ell+1}) \tag{16}$$

It is easy to verify that $\{h_\ell\}_\ell$ is eqicontinuous on X and hence by the Strong Law of Large Numbers, $\{h_\ell\}_\ell$ converges to h uniformly on X. Hence, it follows from (16) that

$$h(x^\infty) = \lim_{\ell(\in\mathcal{L})\to\infty} h_\ell(x^{\ell+1}) - \frac{c}{4}\|x^{\ell+1} - x^\ell\|^2 \leq \lim_{\ell(\in\mathcal{L})\to\infty} \alpha_{t_\ell}^\ell + (\beta_{t_\ell}^\ell)^\top x^{\ell+1} \leq h(x^\infty). \tag{17}$$

Equation (17) implies that $\lim\limits_{\ell(\in\mathcal{L})\to\infty} \alpha_{t_\ell}^\ell + (\beta_{t_\ell}^\ell)^\top x^{\ell+1} = h(x^\infty)$. Since $\mathcal{L}_0 \subseteq \mathcal{L}$, we have

$$h(x^\infty) = \lim_{\ell(\in\mathcal{L}_0)\to\infty} \alpha_{t_\ell}^\ell + (\beta_{t_\ell}^\ell)^\top x^{\ell+1} = \alpha^\infty + (\beta^\infty)^\top x^\infty. \tag{18}$$

On the other hand, for any $x \in X$, $\alpha_{t_\ell}^\ell + (\beta_{t_\ell}^\ell)^\top x \leq \hat{h}_\ell(x) \leq h_\ell(x)$, which implies that

$$\lim_{\ell(\in\mathcal{L}_0)\to\infty} \alpha_{t_\ell}^\ell + (\beta_{t_\ell}^\ell)^\top x = \alpha^\infty + (\beta^\infty)^\top x \leq h(x). \tag{19}$$

Thus, it follows from the combination of equation (17), (18), and (19) that $\beta^\infty \in \partial h(x^\infty)$. Therefore,

$$\begin{aligned}\limsup_{\ell(\in\mathcal{L})\to\infty} h'_\ell(x^{\ell+1}; x - x^{\ell+1}) &= (\beta^\infty)^\top (x - x^\infty) \\ &\leq \max\{u^\top (x - x^\infty) \mid u \in \partial h(x^\infty)\} \\ &= h'(x^\infty; x - x^\infty).\end{aligned} \tag{20}$$

■

The following theorem shows that the directional derivative operator and the expectation operator can be swapped under mild assumptions.

Theorem 1. *If Assumptions A1 - A4 hold, then for any $\hat{x} \in X$, $f(x)$ is directionally differentiable $\hat{x}$ and*

$$f'(\hat{x}; x - \hat{x}) = \mathbb{E}_{\tilde{\xi}}[F'_{\tilde{\xi}}(\hat{x}; x - \hat{x})], \ \forall \ x \in X$$

Furthermore, if Assumption A4 is replaced by "$F(\cdot, \tilde{\xi})$ is differentiable on X for almost every $\tilde{\xi}$", then the following alternative conclusion holds: For any $\hat{x} \in X$, $f(x)$ is differentiable at $\hat{x}$ and

$$\nabla f(\hat{x}) = \mathbb{E}_{\tilde{\xi}}[\nabla_x F(\hat{x}, \tilde{\xi})].$$

Proof. Proof. See Theorem 7.44 (b) and (c) of Shapiro et al. [44]. ■

Theorem 2. *Suppose Assumptions A1 - A6 hold. Suppose either of the following two conditions holds:*

A7 $F(\cdot, \xi)$ is differentiable and concave on X for almost every ξ.

A8 $F(\cdot, \xi)$ is L-smooth on X for almost every ξ.

Then, with probability one, any accumulation point of $\{x^\ell\}_\ell$ is a d-stationary point of (5).

Proof. **Case 1: Assumption A7 holds.**
In this case, the surrogate function of $F(x, \xi)$ near $x' \in X$ is written as: $G(x, \xi; x') = F(x', \xi) + \nabla_x F(x', \xi)^\top (x - x')$ for $x \in X$. Since X is compact and $\{x^\ell\}_\ell \subset X$, the accumulation point of $\{x^\ell\}_\ell$ exists, say $\lim_{\ell(\in\mathcal{L})\to\infty} x^\ell = x^\infty \in X$. By Proposition 3, we also have $\lim_{\ell(\in\mathcal{L})\to\infty} x^{\ell+1} = x^\infty$. By the update rule in Algorithm 1, we have

the optimality condition in the ℓ^{th} iteration as follows: $\forall\, x \in X$,

$$0 \leq (\frac{1}{\ell}\sum_{i=1}^{\ell} \nabla_x F(x^\ell, \xi_i))^\top (x - x^{\ell+1}) + \hat{h}'_{\ell,\nu_\ell}(x^{\ell+1}; x - x^{\ell+1}) + c(x^{\ell+1} - x^\ell)^\top (x - x^{\ell+1}). \tag{21}$$

By Assumption A3 and Strong Law of Large Number (theorem 2.5.6 in Durrett [13]), $\lim\limits_{\ell(\in\mathcal{L})\to\infty} \frac{1}{\ell}\sum_{i=1}^{\ell} \nabla_x F(x^\ell, \xi_i)$ $= \mathbb{E}_{\tilde{\xi}}[\nabla_x F(x^\ell, \tilde{\xi})]$. By Theorem 1, it further implies that $\lim\limits_{\ell(\in\mathcal{L})\to\infty} \frac{1}{\ell}\sum_{i=1}^{\ell} \nabla_x F(x^\infty, \xi_i) = \nabla \mathbb{E}_{\tilde{\xi}}[F(x^\infty, \tilde{\xi})] =$ $\nabla f(x^\infty)$. By Proposition 4, it follows from (23) that

$$\begin{aligned}
0 \leq & \liminf_{\ell(\in\mathcal{L})\to\infty} \{(\frac{1}{\ell}\sum_{i=1}^{\ell} \nabla_x F(x^\ell, \xi_i))^\top (x - x^{\ell+1}) + \hat{h}'_{\ell,\nu_\ell}(x^{\ell+1}; x - x^{\ell+1}) \\
& + c(x^{\ell+1} - x^\ell)^\top (x - x^{\ell+1})\} \\
& \leq \nabla f(x^\infty)^\top (x - x^\infty) + \limsup_{\ell(\in\mathcal{L})\to\infty} \hat{h}'_{\ell,\nu_\ell}(x^{\ell+1}; x - x^{\ell+1}) \\
& \leq \nabla f(x^\infty)^\top (x - x^\infty) + h'(x^\infty; x - x^\infty), \ \forall\, x \in X.
\end{aligned} \tag{22}$$

Case 2: Assumption A8 holds. In this case, the surrogate function of $F(x,\xi)$ near $x' \in X$ is written as (by the descent lemma in [4]): $G(x,\xi;x') = F(x,\xi) + \nabla_x F(x,\xi)^\top (x - x') + \frac{L}{2}\|x - x'\|^2$. Again, the optimality condition in the ℓ^{th} iteration is as follows: $\forall\, x \in X$,

$$0 \leq (\frac{1}{\ell}\sum_{i=1}^{\ell} \nabla_x F(x^\ell, \xi_i))^\top (x - x^{\ell+1}) + \hat{h}'_{\ell,\nu_\ell}(x^{\ell+1}; x - x^{\ell+1}) + (c+L)(x^{\ell+1} - x^\ell)^\top (x - x^{\ell+1}). \tag{23}$$

By following the same procedure sa in Case 1, we can conclude that in (22) holds. ■

3 Nonconvex Predictive Stochastic Programming

Recall in PSP, we consider the random vector appears as a tuple $(\tilde{\omega}, \tilde{\xi}) : \Omega \mapsto \mathcal{X} \times \mathcal{Y}$, where $\tilde{\omega}$ correlates with the previously introduced response, $\tilde{\xi}$. This section first introduces a base MM algorithm for solving nonconvex PSP problems and then combines it with the SD-MM algorithm in the previous section to design its non-parametric extension. We shall refer to the base MM algorithm as N-LEON.

3.1 N-LEON

Before we introduce the problem, we make the following assumption of the relation of objective function $(F : X \times \mathcal{Y} \mapsto \mathbb{R})$ and the tuple $(\tilde{\omega}, \tilde{\xi})$. We do not assume that $F(\cdot, \xi)$ is differentiable on feasible region X

in this subsection.

B1 X is a compact convex set.

B2 $|F(x,\tilde{\xi})|$ is uniformly bounded with probability one (i.e., there exists M such that $|F(x,\tilde{\xi})| \leq M < \infty \; \forall \; x \in X \; w.p.1$).

B3 For almost every $\tilde{\xi}$, $F(x,\tilde{\xi})$ is Lipschitz continuous on X with a common Lipschitz modulus, Lip_F.

B4 $F(\cdot,\tilde{\xi})$ is directionally differentiable on X for almost every $\tilde{\xi}$.

B5 $(\tilde{\omega},\tilde{\xi})$ follows a joint distribution $\mu_{\tilde{\omega},\tilde{\xi}}$. The regular conditional distribution of $\tilde{\xi}$ given ω exists, which is denoted by $\mu_{\tilde{\xi}|\omega}$.

Here, we consider solving a PSP as follows:

$$\min_{x\in X} f_\omega(x) = \mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[F(x,\tilde{\xi})|\tilde{\omega}=\omega]. \tag{24}$$

Given a dataset, $S_n \triangleq \{(\omega_i,\xi_i)\}_{i=1}^n$, where $\{(\omega_i,\xi_i)\}_{i=1}^n$ are n realizations of the i.i.d. copies of $(\tilde{\omega},\tilde{\xi}) \sim \mu_{\tilde{\omega},\tilde{\xi}}$. We define the k-nearest neighbors estimate of $f_\omega(x)$ as

$$f_{\omega,n}^{k_n}(x) = \frac{1}{k_n}\sum_{i=1}^n \mathbb{I}(\omega_i \in \mathcal{S}(k_n,\omega;\{\omega_i\}_{i=1}^n))F(x,\xi_i). \tag{25}$$

First introduced by Fix and Hodges in [15], kNN method has been widely used in hedonic housing pricing (Oladunni and Sharma [36]), traffic flow prediction (Smith and Demetsky [45]), battery capacity prediction (Hu et al. [19]), wind power forecast (Mangalova and Agafonov [34]) and so on. On the other hand, the asymptotic properties of kNN method has been extensively studied by Stone [46], Devroye et al. [11], Györfi et al.[16], Walk [48] and so on. Now we introduce the pointwise convergence of the kNN estimator below, which is fundamental to the convergence of the following proposed algorithms for PSP.

Theorem 3. *Suppose that assumptions B1 - B5 are satisfied. Further suppose* $\{(\omega_i,\xi_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu_{\tilde{\omega},\tilde{\xi}}$. *Let* k_n *be monotonically increasing with* n, $k_\ell \to \infty$, $\frac{k_n}{n} \to 0$ *(as* $\ell \to \infty$*) and* (k_n) *varies regularly with exponent*

$\beta \in (0,1]$ (e.g., $k_n = \lfloor n^{\beta} \rfloor$, $\beta \in (0,1)$). Then the following holds:

$$\lim_{n\to\infty} \frac{1}{k_n} \sum_{i=1}^{n} \mathbb{I}\left(\tilde{\omega}_i \in \mathcal{S}(k_n, \omega; \{\tilde{\omega}_j\}_{j=1}^{n})\right) F(x, \tilde{\xi}_i) = \mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}}[F(x, \tilde{\xi}) | \tilde{\omega} = \omega], \quad w.p.1.$$

Proof. See theorem 1 of Walk [48]. ■

One key benefit of using kNN is that $\frac{1}{k_n} \sum_{i=1}^{n} \mathbb{I}(\omega_i \in \mathcal{S}(k_n, \omega; \{\omega_i\}_{i=1}^{n})) = 1$ and $\frac{1}{k_n} \mathbb{I}(\omega_i \in \mathcal{S}(k_n, \omega; \{\omega_i\}_{i=1}^{n})) \geq 0$ for all $i \in \{1, 2, \ldots, n\}$, which implies that many results hold for the sample averaging can also be transferred to the algorithm powered by kNN. Note that $f_{\omega,n}^{k_n}(x)$ is the biased estimate of $f_\omega(x)$ and it is unclear whether $\mathbb{E}_{\tilde{S}_n}[\|\nabla_x f_{\omega,n}^{k_n}(x) - \nabla_x f_\omega(x)\|^2]$ can be bounded by some function that depends on the sample size n. Although Györfi et al. [16] in Theorem 6.2 derives the rate of convergence of kNN estimation in terms of $\int_\omega |f_{\omega,n}^{k_n}(x) - \mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}}[F(x, \tilde{\xi}) | \tilde{\omega} = \omega]| \mu_\omega(d_\omega)$ for a given x, the rate of convergence of $|f_{\omega,n}^{k_n}(x) - \mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}}[F(x, \tilde{\xi}) | \tilde{\omega} = \omega]|$ is still unknown. To overcome this challenge, we use the incumbent selection rule based on the regularized SD in Higle and Sen [18] to recover a sequence of convergent incumbents.

Algorithm 4 N-LEON-kNN

```
(Initialization) Pick $\bar{x}^1 \in X$, $c > 0$, $r \in (0,1)$, and $\beta \in (0,1)$. Let $\ell = 1$. Let $k_\ell = \lfloor \ell^{\beta} \rfloor$. Generate $(\omega_1, \xi_1) \sim \mu_{\tilde{\omega},\tilde{\xi}}$.
(Step 1: Candidate Selection)
    $x^{\ell+1} = \arg\min_{x \in X} \frac{1}{k_\ell} \sum_{i=1}^{\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell, \omega; \{\omega_i\}_{i=1}^{\ell})) G(x, \xi_i; \bar{x}^\ell) + \frac{c}{2}\|x - \bar{x}^\ell\|^2.$
(Step 2) Generate $(\omega_{\ell+1}, \xi_{\ell+1}) \sim \mu_{\tilde{\omega},\tilde{\xi}}$, which is independent from the past samples.
(Step 3: Incumbent Selection)
if $f_{\omega,\ell+1}^{k_{\ell+1}}(x^{\ell+1},) - f_{\omega,\ell+1}^{k_{\ell+1}}(\bar{x}^\ell) < r\left[f_{\omega,\ell}^{k_\ell}(x^{\ell+1}) - f_{\omega,\ell}^{k_\ell}(\bar{x}^\ell)\right]$ then
    Set $\bar{x}^{\ell+1} = x^{\ell+1}$
else
    Set $\bar{x}^{\ell+1} = \bar{x}^\ell$
end if
(Step 4) Set $\ell \leftarrow \ell + 1$ and go to Step 1.
```

Proposition 5. *$\{x^\ell\}_\ell$ and $\{\bar{x}^\ell\}_\ell$ generated by Algorithm 4 satisfies the following relation:*

$$f_{\omega,\ell}^{k_\ell}(x^{\ell+1}) - f_{\omega,\ell}^{k_\ell}(\bar{x}^\ell) \leq -\frac{c}{2}\|x^{\ell+1} - \bar{x}^\ell\|^2. \tag{26}$$

Proof. By the optimality of Candidate Selection in Algorithm 4, we have

$$\begin{aligned}&\frac{1}{k_\ell}\sum_{i=1}^{\ell}\mathbb{I}(\omega_i \in \mathcal{S}(k_\ell,\omega;\{\omega_i\}_{i=1}^{\ell}))G(x,\xi_i;\bar{x}^\ell) + \frac{c}{2}\|x^{\ell+1}-\bar{x}^\ell\|^2\\ &\le \frac{1}{k_\ell}\sum_{i=1}^{\ell}\mathbb{I}(\omega_i \in \mathcal{S}(k_\ell,\omega;\{\omega_i\}_{i=1}^{\ell}))G(\bar{x}^\ell,\xi_i;\bar{x}^\ell).\end{aligned} \tag{27}$$

The final result follow from the definition of the surrogate function in Definition 2. ■

Proposition 6. *Suppose Assumption B1 - B5 hold. Then* $\lim_{n\to\infty} f^{k_n}_{\omega,n}(x) = f_\omega(x)$ *uniformly on* X *with probability one.*

Proof. Assumption B3 implies that $\{f^{k_n}_{\omega,n}(x)\}_n$ is Lipschitz continuous with a common modulus, Lip_F, which further implies the equicontinuity of $\{f^{k_n}_{\omega,n}(x)\}_n$. By Theorem 3, $\{f^{k_n}_{\omega,n}(x)\}_n$ converges pointwise to $\mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[F(x,\tilde{\xi})|\tilde{\omega}=\omega]$ on X with probability one. Finally, the equicontinuity and pointwise convergence of $\{f^{k_n}_{\omega,n}(x)\}_n$ implies the uniform convergence. ■

Inspired by Lemma 6 in [17], we show a key limiting property of the estimated function value at the candidates and incumbents.

Proposition 7. *Suppose Assumption B1 - B5 hold. Then*

$$\limsup_{\ell\to\infty} f^{k_\ell}_{\omega,\ell}(x^{\ell+1}) - f^{k_\ell}_{\omega,\ell}(\bar{x}^\ell) = 0 \ w.p.1. \tag{28}$$

Proof. Proof. Case 1: $\{\bar{x}^\ell\}$ changes finitely often. Then there exists $\ell > 0$ and $\bar{x} \in X$, such that $\bar{x}^\ell = \bar{x}$ for all $\ell \geq L$. Then the incumbent selection step implies that

$$f^{k_\ell}_{\omega,\ell}(x^\ell) - f^{k_\ell}_{\omega,\ell}(\bar{x}) \geq r[f^{k_{\ell-1}}_{\omega,\ell-1}(x^\ell) - f^{k_{\ell-1}}_{\omega,\ell-1}(\bar{x})] \ \forall \ \ell \geq L+1. \tag{29}$$

Since X is compact, the accumulation point of $\{x^\ell\}$ exists. Let $\{x^\ell\}_{\ell\in\mathcal{L}}$ be a subsequence whose accumulation point is $\hat{x}$. By Proposition 6, we have $\lim_{\ell(\in\mathcal{L})\to\infty} f^{k_\ell}_{\omega,\ell}(x^\ell) = \lim_{\ell(\in\mathcal{L})\to\infty} f^{k_\ell}_{\omega,\ell-1}(x^\ell) = f_\omega(\hat{x})$ and $\lim_{\ell(\in\mathcal{L})\to\infty} f^{k_{\ell-1}}_{\omega,\ell-1}(\bar{x}) = \lim_{\ell(\in\mathcal{L})\to\infty} f^{k_{\ell-1}}_{\omega,\ell}(\bar{x}) = f_\omega(\bar{x})$. Let $\ell(\in\mathcal{L})\to\infty$, equation (29) implies that

$$f_\omega(\hat{x}) - f_\omega(\bar{x}) \geq r[f_\omega(\hat{x}) - f_\omega(\bar{x})]. \tag{30}$$

Since $f_{\omega,\ell-1}(x^\ell) - f_{\omega,\ell-1}(\bar{x}) \leq 0$ by Proposition 5, we have $\limsup_{\ell\to\infty}[f_{\omega,\ell-1}(x^\ell) - f_{\omega,\ell-1}(\bar{x})] \leq 0$, which implies that $f_\omega(\hat{x}) - f_\omega(\bar{x}) = \lim_{\ell(\in\mathcal{L})\to\infty} [f_{\omega,\ell-1}(x^\ell) - f_{\omega,\ell-1}(\bar{x})] \leq 0$. Thus, combined equation (30), we have $f_\omega(\hat{x}) - f_\omega(\bar{x}) = 0$ which completes the proof in Case 1.

Case 2: $\{\bar{x}^\ell\}$ changes infinitely often. Let $\{\ell_n\}$ denote the subsequence of iterations so that $\{\bar{x}^{\ell_n}\}$ changes. By the incumbent selection, we have

$$f^{k_{\ell_n}}_{\omega,\ell_n}(x^{\ell_n}) - f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n-1}) \leq r\left[f^{k_{\ell_n-1}}_{\omega,\ell_n-1}(x^{\ell_n}) - f^{k_{\ell_n-1}}_{\omega,\ell_n-1}(\bar{x}^{\ell_n-1})\right] \leq 0. \tag{31}$$

Let $\theta^{\ell_n} = f^{k_{\ell_n-1}}_{\omega,\ell_n-1}(x^{\ell_n}) - f^{k_{\ell_n-1}}_{\omega,\ell_n-1}(\bar{x}^{\ell_n-1})$. Then for $m > 0$

$$\frac{1}{m}\sum_{n=1}^{m}[f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n}) - f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n-1})] \leq \frac{r}{m}\sum_{n=1}^{m}\theta^{\ell_n} \leq 0. \tag{32}$$

Also note that

$$\begin{aligned}
&\frac{1}{m}\sum_{n=1}^{m}[f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n}) - f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n-1})] \\
&= \frac{1}{m}\left(\sum_{n=1}^{m-1}(f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n}) - f^{k_{\ell_{n+1}}}_{\omega,\ell_{n+1}}(\bar{x}^{\ell_n})) + f^{k_{\ell_m}}_{\omega,\ell_m}(\bar{x}^{\ell_m}) - f^{k_{\ell_1}}_{\omega,\ell_1}(\bar{x}^{\ell_0})\right) \\
&= \frac{1}{m}\sum_{n=1}^{m-1}\left(f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n}) - f^{k_{\ell_{n+1}}}_{\omega,\ell_{n+1}}(\bar{x}^{\ell_n})\right) + \frac{1}{m}(f^{k_{\ell_m}}_{\omega,\ell_m}(\bar{x}^{\ell_m}) - f^{k_{\ell_1}}_{\omega,\ell_1}(\bar{x}^{\ell_0})) \\
&\geq \frac{1}{m}\sum_{n=1}^{m-1}\left(f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n}) - f^{k_{\ell_{n+1}}}_{\omega,\ell_{n+1}}(\bar{x}^{\ell_n})\right) - \frac{2M}{m}.
\end{aligned} \tag{33}$$

The last inequality holds by Assumption B2. By Proposition 6, $\lim_{n\to\infty}(f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n}) - f^{k_{\ell_{n+1}}}_{\omega,\ell_{n+1}}(\bar{x}^{\ell_n})) = 0$ with probability one. Thus, let $m \to \infty$, it follows from (33) that $\lim_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m}[f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n}) - f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n-1})] = 0$ *w.p.*1. Thus, (32) implies that

$$0 = \lim_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m}[f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n}) - f^{k_{\ell_n}}_{\omega,\ell_n}(\bar{x}^{\ell_n-1})] \leq \lim_{m\to\infty}\frac{r}{m}\sum_{n=1}^{m}\theta^{\ell_n} \leq 0 \ w.p.1 \tag{34}$$

Equation (34) implies that $0 = \limsup_{m\to\infty}\frac{1}{m}\sum_{n=1}^{m}\theta^{\ell_n} \leq \limsup_{m\to\infty}\theta^{\ell_m} \leq 0$ *w.p.*1, which completes the proof. ■

Analogous to Theorem 1 (also can be seen in Theorem 7.44 in [44]), we shall get the kNN version of the results by using the "strong law of large numbers" of kNN estimators (Theorem 3).

Theorem 4. *If Assumptions B1 - B5 hold, then for any $\hat{x} \in X$, $f(x,\omega)$ is directionally differentiable at $\hat{x}$*

and

$$f'_\omega(\hat{x}; x-\hat{x}) = \mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[F'_{\tilde{\xi}}(\hat{x}; x-\hat{x})|\tilde{\omega}=\omega],\ \forall\ x\in X \tag{35}$$

Furthermore, if Assumption A4 is replaced by "$F(\cdot,\tilde{\xi})$ is differentiable on X for almost every $\tilde{\xi}$", then the following alternative conclusion holds: For any $\hat{x}\in X$, $f(x)$ is differentiable at $\hat{x}$ and

$$\nabla_x f_\omega(\hat{x}) = \mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[\nabla_x F(\hat{x},\tilde{\xi})|\tilde{\omega}=\omega]. \tag{36}$$

Proof. The proof is similar to the proof of Theorem 7.44 (b) and (c) of [44]. The only difference is that we use regular conditional distribution to take the integral when performing Dominated Convergence Theorem. ■

With the Propositions 5 and 7 providing a subsequence so that distance between the incumbents reduce to zero, we can show convergence of N-LEON-kNN in the following two propositions by providing different structural assumptions on the objective function.

The use of the difference between surrogate function and the objective function in the proposition below is inspired by [32].

Proposition 8. *Let $H(x,\xi;x') \triangleq G(x,\xi;x') - F(x,\xi)$. Suppose Assumptions A1 - A4 hold. Suppose that $H(x,\xi;x^\ell)$ is L-smooth and $\nabla_x H(x',\xi;x') = 0$ for all $x'\in X$ and $\xi\in\Xi$. Then, with probability one, Algorithm 4 produces a subsequence of $\{\bar{x}^\ell\}$ so that $\lim\limits_{\ell(\in\mathcal{L})\to\infty} \|x^{\ell+1}-\bar{x}^\ell\| = 0$ and any accumulation point of such subsequence is a d-stationary point of (24).*

Proof. Throughout this proof, we let $G'_{\xi;x'}(x;d)$ denote the directional derivative of $G(\cdot,\xi;x')$ at x in the direction of d. By Proposition 7, there exists $\mathcal{L}$ such that $\lim_{\ell(\in\mathcal{L})\to\infty}\|x^{\ell+1}-\bar{x}^\ell\| = 0$.
On the other hand, by the optimality condition of (2), for any $x\in X$, we have

$$0 \le \frac{1}{k_\ell}\sum_{i=1}^{\ell}\mathbb{I}(\omega_i\in\mathcal{S}(k_\ell,\omega;\{\omega_i\}_{i=1}^{\ell}))[G'_{\xi_i;\bar{x}^\ell}(x^{\ell+1};x-x^{\ell+1})] + c(x^{\ell+1}-\bar{x}^\ell)^\top(x-x^{\ell+1}). \tag{37}$$

By using

$$F'_{\xi_i}(x^{\ell+1};x-x^{\ell+1}) = G'_{\xi_i;\bar{x}^\ell}(x^{\ell+1};x-x^{\ell+1}) - \nabla_x H(x^{\ell+1},\xi_i;\bar{x}^\ell)(x^{\ell+1};x-x^{\ell+1})$$

and L-smoothness of H, we get

$$\frac{1}{k_\ell}\sum_{i=1}^{\ell}\mathbb{I}(\omega_i\in\mathcal{S}(k_\ell,\omega;\{\omega_i\}_{i=1}^{\ell}))[F'_{\xi_i}(x^{\ell+1};x-x^{\ell+1})] \ge -(c+L)\|x^{\ell+1}-\bar{x}^\ell\|\|x-x^{\ell+1}\| \tag{38}$$

For $\mathcal{L}' \subset \mathcal{L}$ such that $\lim_{\ell(\in \mathcal{L}')\to\infty} \bar{x}^\ell = \bar{x}^\infty$, let $\ell(\in \mathcal{L}') \to \infty$, by the pointwise convergence of kNN estimator (Theorem 3), we have

$$\begin{aligned}
&\lim_{\ell(\in \mathcal{L}')\to\infty} \frac{1}{k_\ell} \sum_{i=1}^{\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell, \omega; \{\omega_i\}_{i=1}^{\ell}))[F'_{\xi_i}(x^{\ell+1}; x - x^{\ell+1})] \\
&= \mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[F'_{\tilde{\xi}}(\bar{x}^\infty; x - \bar{x}^\infty)|\tilde{\omega} = \omega].
\end{aligned}$$

By Theorem 4, it follows that $\mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[F'_{\tilde{\xi}}(\bar{x}^\infty; x - \bar{x}^\infty)|\tilde{\omega} = \omega] = f'_\omega(\bar{x}^\infty, x - \bar{x}^\infty)$. Hence, it follows from (38) that

$$\begin{aligned}
f'_\omega(\bar{x}^\infty, x - \bar{x}^\infty) &= \lim_{\ell(\in \mathcal{L}')\to\infty} \frac{1}{k_\ell} \sum_{i=1}^{\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell, \omega; \{\omega_i\}_{i=1}^{\ell}))[F'_{\xi_i}(x^{\ell+1}; x - x^{\ell+1})] \\
&\geq \lim_{\ell(\in \mathcal{L}')\to\infty} -(c + L)\|x^{\ell+1} - \bar{x}^\ell\|\|x - x^{\ell+1}\| \\
&= 0.
\end{aligned} \tag{39}$$

■

In the following proposition, we consider the case when F is a dc function.

Proposition 9. *Suppose $F(x, \xi) = \psi_1(x, \xi) - \psi_2(x, \xi)$, where $\psi_1 : S \times \mathcal{Y} \mapsto \mathbb{R}$ and $\psi_2 : S \times \mathcal{Y} \mapsto \mathbb{R}$ are convex on a open set $S \subseteq \mathbb{R}^n$ and S is a superset of X. Further assume that Assumptions B1 - B5 hold. Further suppose that $\psi_2(\cdot, \xi)$ is differentiable on S for almost every $\xi \in \mathcal{Y}$.Then, with probability one, Algorithm 4 produces a subsequence of $\{\bar{x}^\ell\}$ so that $\lim_{\ell(\in\mathcal{L})\to\infty} \|x^{\ell+1} - \bar{x}^\ell\| = 0$ and any accumulation point of such a subsequence is a d-stationary point of (24).*

Proof. The surrogate function is

$$G(x, \xi; x') = \psi_1(x, \xi) - [\psi_2(x', \xi) + \nabla_x \psi_2(x', \xi)^\top (x - x')]. \tag{40}$$

Let $\phi_{\omega,1}(x) = \mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[\psi_1(x, \tilde{\xi})|\tilde{\omega} = \omega]$ and $\phi_{\omega,2}(x) = \mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[\psi_2(x, \tilde{\xi})|\tilde{\omega} = \omega]$. Similar to the argument in Proposition 8, there exists $\mathcal{L}$ such that $\lim_{l(\in\mathcal{L})\to\infty} \|x^{\ell+1} - \bar{x}^\ell\| = 0$. In the ℓ^{th} iteration, the optimality condition of **Candidate Selection** in the Algorithm 4 implies that for any $x \in X$, we have

$$\begin{aligned}
0 &\leq c(x^{\ell+1} - \bar{x}^\ell)^\top (x - x^{\ell+1}) \\
&+ \frac{1}{k_\ell} \sum_{i=1}^{\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell, \omega; \{\omega_i\}_{i=1}^{\ell})) \Big[\psi'_{\xi_i,1}(x^{\ell+1}; x - x^{\ell+1}) - \nabla_x \psi_2(\bar{x}^\ell, \xi_i)^\top (x - x^{\ell+1})\Big]
\end{aligned} \tag{41}$$

For $\mathcal{L}' \subset \mathcal{L}$ such that $\lim_{\ell(\in \mathcal{L}') \to \infty} \bar{x}^\ell = \bar{x}^\infty$, let $\ell(\in \mathcal{L}') \to \infty$, by the pointwise convergence of kNN estimator (Theorem 3), we have

$$\begin{aligned}
&\lim_{l(\in \mathcal{L}') \to \infty} \frac{1}{k_\ell} \sum_{i=1}^{\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell, \omega; \{\omega_i\}_{i=1}^{\ell})) \left[\psi_{\xi_i,1'}(x^{\ell+1}; x - x^{\ell+1})\right] \\
&= \mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}} \left[\psi_{\tilde{\xi},1'}(\bar{x}^\infty; x - \bar{x}^\infty) | \tilde{\omega} = \omega\right],
\end{aligned}$$

and $\lim_{l(\in \mathcal{L}') \to \infty} \frac{1}{k_\ell} \sum_{i=1}^{\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell, \omega; \{\omega_i\}_{i=1}^{\ell})) \nabla_x \psi_2(\bar{x}^\ell, \xi_i) = \mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}}[\nabla_x \psi_2(\bar{x}^\infty, \tilde{\xi}) | \tilde{\omega} = \omega]$. It follows from (41) that

$$0 \leq \mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}} \left[\psi_{\tilde{\xi},1'}(\bar{x}^\infty; x - \bar{x}^\infty) | \tilde{\omega} = \omega\right] - \mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}}[\nabla_x \psi_2(\bar{x}^\infty, \tilde{\xi}) | \tilde{\omega} = \omega]. \tag{42}$$

By Theorem 4, we have

$$\mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}} \left[\psi_{\tilde{\xi},1'}(\bar{x}^\infty; x - \bar{x}^\infty) | \tilde{\omega} = \omega\right] = \phi'_{\omega,1}(\bar{x}^\infty; x - \bar{x}^\infty), \tag{43a}$$

$$\mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}}[\nabla_x \psi_2(\bar{x}^\infty, \tilde{\xi})] = \nabla_x \phi_{\omega,2}(\bar{x}^\infty) \tag{43b}$$

Thus, the combination of equations (42) and (43) implies that

$$0 \leq \phi'_{\omega,1}(\bar{x}^\infty; x - \bar{x}^\infty) - \nabla_x \phi_{\omega,2}(\bar{x}^\infty)^\top (x - \bar{x}^\infty). \tag{44}$$

Thus, this shows that $\bar{x}^\infty$ is the d-stationary point of (24). ■

3.2 NSD-MM

In this subsection, we aim to solve the predictive version of (5) in section 2.

$$\min_{x \in X} \mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}} \left[F(x, \tilde{\xi}) + H(x, \tilde{\xi}) | \tilde{\omega} = \omega\right] \tag{45}$$

Same as the objective function setup in section 2, we let $F : S \times \mathcal{Y} \mapsto \mathbb{R}$ be a L-smooth/differentiable concave function, where S is a superset of X, and let $H(x, \xi)$ be a pointwise maximum of finitely many linear functions of x. To the end of this subsection, we shall assume that Assumptions A1 - A6 and B5 hold. Additionally, we assume that

B6 There exists $M_h \in (0, \infty)$ such that $H(x, \xi) < M_h$ for all $x \in X$ and almost every $\xi \in \mathcal{Y}$.

Assumption B6 is important in updating the piecewise linear approximation function of the convex component when kNN being used. Throughout this subsection, we make the following notations:

$$
\begin{aligned}
&\ell: \text{ iteration number of the outer loop, } \nu: \text{ iteration number of the inner loop}\\
&\omega: \text{ observed predictor, } N_\ell: \text{ sample size in the } \ell^{\text{th}} \text{ outer loop}\\
&\nu_\ell: \text{ number of inner loops in the outer iteration } \ell\\
&\{\bar{x}^\ell\}: \text{ sequence of incumbents generated by the outer loop}\\
&\{x_\nu^{\frac{1}{2},\ell}\}: \text{ sequence of candidates generated by the inner loops}\\
&\{x^\ell\}: \text{ sequence of candidates generated by the outer loops}\\
&f_{\omega,\ell}^{k_\ell}(x) = \frac{1}{k_\ell}\sum_{i=1}^{N_\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell, \omega; \{\omega_i\}_{i=1}^{N_\ell}))F(x,\xi_i)\\
&h_{\omega,\ell}^{k_\ell}(x) = \frac{1}{k_\ell}\sum_{i=1}^{N_\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell, \omega; \{\omega_i\}_{i=1}^{N_\ell}))H(x,\xi_i)\\
&\zeta_{\omega,\ell}^{k_\ell}(x) = f_{\omega,\ell}^{k_\ell}(x) + h_{\omega,\ell}^{k_\ell}(x)\\
&f_\omega(x) = \mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[F(x,\tilde{\xi})|\tilde{\omega}=\omega], \; h_\omega(x) = \mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[H(x,\tilde{\xi})|\tilde{\omega}=\omega]\\
&\zeta_\omega(x) = f_\omega(x) + h_\omega(x)\\
&\hat{h}_{\omega,\ell,\nu}^{k_\ell}(x): \text{ piecewise linear approximation of } h_{\omega,\ell}^{k_\ell} \text{ in the } \nu^{\text{th}} \text{ inner iteration of the } \ell^{\text{th}} \text{ outer loop}\\
&g_{\omega,\ell}^{k_\ell}(x;x') = \frac{1}{k_\ell}\sum_{i=1}^{\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell, \omega; \{\omega_i\}_{i=1}^{\ell}))G(x,\xi_i;x')
\end{aligned}
$$

For all $i \in \{1,\ldots,N_{\ell+1}\}$, compute $\bar{\pi}_i^\ell \in \arg\max\{\pi^\top(e(\xi_i)-C(\xi_i)\bar{x}^\ell) : \pi^\top D \leq d\}$ and $\pi_i^{\ell+1} \in \arg\max\{\pi^\top(e(\xi_i)-C(\xi_i)x^{\ell+1}) : \pi^\top D \leq d\}$ for $i = 1,2,\ldots,N_\ell$. The new minorants for the $(\ell+1)^{st}$ iteration are constructed by using the following formula.

$$
\begin{aligned}
u_{\omega,\ell+1}^{k_{\ell+1}}(\bar{x}^\ell) &= \frac{1}{k_{\ell+1}}\sum_{i=1}^{N_{\ell+1}} \mathbb{I}(\omega_i \in \mathcal{S}(k_{\ell+1}, \omega; \{\omega_i\}_{i=1}^{N_{\ell+1}}))C(\xi_i)^\top \bar{\pi}_i^\ell,\\
\mathcal{C}_{\omega,\ell+1}^1(x) &= h_{\omega,\ell+1}^{k_{\ell+1}}(\bar{x}^\ell) + u_{\omega,\ell+1}^{k_{\ell+1}}(\bar{x}^\ell)^\top(x-\bar{x}^\ell)\\
u_{\omega,\ell+1}^{k_{\ell+1}}(x^{\ell+1}) &= \frac{1}{k_{\ell+1}}\sum_{i=1}^{N_{\ell+1}} \mathbb{I}(\omega_i \in \mathcal{S}(k_{\ell+1}, \omega; \{\omega_i\}_{i=1}^{N_{\ell+1}}))C(\xi_i)^\top \pi_i^{\ell+1}\\
\mathcal{C}_{\omega,\ell+1}^2(x) &= h_{\omega,\ell+1}^{k_{\ell+1}}(x^{\ell+1}) + u_{\omega,\ell+1}^{k_{\ell+1}}(x^{\ell+1})^\top(x-x^{\ell+1})
\end{aligned}
\tag{46}
$$

The algorithm design can be regarded as a mix of SD-MM and N-LEON-kNN. In the inner loop, we still require the lower-bounding approximation must be accurate enough (i.e., $h_{\ell}^{k_\ell}(x_{\nu}^{\frac{1}{2},\ell+1},\omega) - \hat{h}_{\ell}^{k_\ell}(x_{\nu}^{\frac{1}{2},\ell+1},\omega) \leq \frac{c}{4}\|x_{\nu}^{\frac{1}{2},\ell+1} - \bar{x}^{\ell}\|^2$). We also introduce an incumbent selection similar to the one in Algorithm 4 to ensure it produces a subsequence of incumbents so that any accumulation point on the subsequence is a d-stationary point of (3). It is also worth noting the convergence result of Algorithm 5 is weaker than the one of Algorithm 1 due to the unknown conditional distribution and the unknown rate of convergence of the non-parametric estimation.

Algorithm 5 NSD-MM

```
(Initialization) Pick $\bar{x}^1 \in X$, $N_1 > 0$, $c > 0$, $\beta \in (0,1)$, and $L > 0$ (number of iterations). Set $\nu = 0$, and $k_1 = \lfloor N_1^{\beta} \rfloor$. Generate i.i.d. $(\omega_i, \xi_i) \sim \mu_{\tilde{\omega},\tilde{\xi}}$ for all $i \in \{1, 2, \ldots, N_1\}$. Calculate $\beta_{\hat{j}}(\xi_i)$ for some $j \in \mathcal{M}(\bar{x}^1, \xi_i)$ and for all $i \in \{1, \ldots, N_1\}$ and calculate $u_{\omega,1}^{1}(\bar{x}^1) = \frac{1}{k_1}\sum_{i=1}^{N_1} \mathbb{I}(\omega_i \in \mathcal{S}(k_{\ell+1}, \omega; \{\omega_i\}_{i=1}^{N_\ell}))\beta_{\hat{j}}(\xi_i)$. Set $\hat{h}_{\omega,1,1}^{k_1}(x) = u_{\omega,1}^{1}(\bar{x}^1)^\top (x - \bar{x}^1)$.
for $\ell = 1, 2, \ldots, L$ do                                                    Outer Loop
    do                                                                              Inner Loop
        Set $\nu \leftarrow \nu + 1$.
        (Candidate Update) $x_{\nu}^{\frac{1}{2},\ell+1} = \arg\min_{x \in X} g_{\omega,\ell}^{k_\ell}(x; \bar{x}^{\ell}) + \hat{h}_{\omega,\ell,\nu}^{k_\ell}(x) + \frac{c}{2}\|x - \bar{x}^{\ell}\|^2$.
        Compute $\pi_{\nu,i}^{\frac{1}{2},\ell+1} \in \arg\max\{\pi^\top (e(\xi_i) - C(\xi_i) x_{\nu}^{\frac{1}{2},\ell+1}) : \pi^\top D \leq d\}$ for $i = 1, 2, \ldots, N_\ell$
        Compute $u_{\omega,\ell}(x_{\nu}^{\frac{1}{2},\ell+1}) = \frac{1}{k_\ell}\sum_{i=1}^{N_\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell, \omega; \{\omega_i\}_{i=1}^{N_\ell}))C(\xi_i)^\top \pi_{\nu,i}^{\frac{1}{2},\ell+1}$.
        Set $\underbrace{\hat{h}_{\omega,\ell,\nu+1}^{k_\ell}(x) \leftarrow \max\{\hat{h}_{\omega,\ell,\nu}^{k_\ell}(x), h_{\omega,\ell}(x_{\nu}^{\frac{1}{2},\ell+1}) + u_{\omega,\ell}(x_{\nu}^{\frac{1}{2},\ell+1})^\top (x - x_{\nu}^{\frac{1}{2},\ell+1})\}}_{\text{Piecewise Linear Approximation Update III}}$.
    while $h_{\omega,\ell}^{k_\ell}(x_{\nu}^{\frac{1}{2},\ell+1}) - \hat{h}_{\omega,\ell}^{k_\ell}(x_{\nu}^{\frac{1}{2},\ell+1}) > \frac{c}{4}\|x_{\nu}^{\frac{1}{2},\ell+1} - \bar{x}^{\ell}\|^2$          End Inner Loop
    Set $x^{\ell+1} \leftarrow x_{\nu}^{\frac{1}{2},\ell+1}$, $N_{\ell+1} \leftarrow N_\ell + 1$
    Generate $(\omega_{N_{\ell+1}}, \xi_{N_{\ell+1}}) \sim \mu_{\tilde{\omega},\tilde{\xi}}$, which is independent from the past samples.
    Perform Minorant Pruning I/II for $\hat{h}_{\omega,\ell,\nu+1}^{k_\ell}(x)$.
    Calculate $k_{\ell+1} = \lfloor N_{\ell+1}^{\beta} \rfloor$ and $\mathcal{C}_{\omega,\ell+1}^{1}(x)$ and $\mathcal{C}_{\omega,\ell+1}^{2}(x)$ using (46).
    if $k_{\ell+1} = k_\ell$ then
        Set $\underbrace{\hat{h}_{\omega,\ell+1,1}^{k_{\ell+1}}(x) \leftarrow \max\{\hat{h}_{\omega,\ell,\nu+1}^{k_\ell}(x) - \frac{M_h}{k_\ell}, \mathcal{C}_{\omega,\ell+1}^{1}(x), \mathcal{C}_{\omega,\ell+1}^{2}(x)\}}_{\text{Piecewise Linear Approximation Update I}}$
    else
        Set $\underbrace{\hat{h}_{\omega,\ell+1,1}^{k_{\ell+1}}(x) \leftarrow \max\{\frac{k_\ell}{k_{\ell+1}}\hat{h}_{\omega,\ell,\nu+1}^{k_\ell}(x), \mathcal{C}_{\omega,\ell+1}^{1}(x), \mathcal{C}_{\omega,\ell+1}^{2}(x)\}}_{\text{Piecewise Linear Approximation Update II}}$
    end if
    if $\zeta_{\omega,\ell+1}^{k_{\ell+1}}(x^{\ell+1}) - \zeta_{\omega,\ell+1}^{k_{\ell+1}}(\bar{x}^{\ell}) < r\left[\zeta_{\omega,\ell}^{k_\ell}(x^{\ell+1}) - \zeta_{\omega,\ell}^{k_\ell}(\bar{x}^{\ell})\right]$ then          Incumbent Selection
        Set $\bar{x}^{\ell+1} \leftarrow x^{\ell+1}$
    else
        Set $\bar{x}^{\ell+1} \leftarrow \bar{x}^{\ell}$
    end if
    Set $\nu_\ell \leftarrow \nu$ and $\nu \leftarrow 0$.
end for                                                                             End Outer Loop
```

The proposition below shows that the combination of **Piecewise Linear Approximation Update I** and **Piecewise Linear Approximation Update II** will provide lower bounding approximation of $h_{\omega,\ell}^{k_\ell}$ for all ℓ.

Proposition 10. *Suppose that Assumptions A1 - A6 and B5 - B6 hold. Tie-breaking by indices is used. With probability one,* $\hat{h}^{k_\ell}_{\omega,\ell,\nu}(x) \leq h^{k_\ell}_{\omega,\ell}(x)\ \forall\ x \in X$ *for all* $\ell \geq 1,\ \nu \geq 1$.

Proof. Let $\omega^\ell_{[k_\ell]}$ be the k_ℓ nearest neighbor of ω from the set $\{\omega_j\}_{j=1}^{N_\ell}$ and let $\xi^\ell_{[k_\ell]}$ be the associated response of $\omega^\ell_{[k_\ell]}$. We let $\hat{h}^{k_1}_{\omega,1,\nu}$ denote the updated piecewise linear approximation function after performing the ν^{th} **Piecewise Linear Approximation Update III** in the ℓ^{th} outer loop.

By the initilization of Algorithm 5, $\hat{h}^{k_1}_{\omega,1,1}(x) = \mathcal{C}^1_{\omega,1}(x) \leq h^{k_1}_{\omega,1}(x),\ \forall\ x \in X$.

Case I: $\ell = 1$ and the inner loop is terminated. By the convexity of $h^{k_2}_{\omega,2}(\cdot)$ and (46), $h^{k_2}_{\omega,2}(x) \geq \mathcal{C}^1_{\omega,2}(x)$ and $h^{k_2}_{\omega,2}(x) \geq \mathcal{C}^2_{\omega,2}(x)$ for all $x \in X$. If $k_2 = k_1$ and $\|\omega^1_{[k_1]} - \omega\| \leq \|\omega_{N_2} - \omega\|$, then $\mathbb{I}(\omega_{N_2} \in \mathcal{S}(k_2, \omega; \{\omega_j\}_{j=1}^{N_2})) = 0$ and $\mathbb{I}(\omega^1_{[i]} \in \mathcal{S}(k_2, \omega; \{\omega_j\}_{j=1}^{N_2})) = 1$ for $i = 1, 2, ..., k_1$. Hence, by the definition of $h^{k_\ell}_\ell$, $h^{k_2}_{\omega,2}(x) = h^{k_2}_{\omega,1}(x)$, which is obvious that $h^{k_2}_{\omega,2}(x) \geq h^{k_2}_{\omega,1}(x) - \frac{M_h}{k_1}\ \forall\ x \in X$. On other hand, if $k_2 = k_1$ and $\|\omega^1_{[k_1]} - \omega\| > \|\omega_{N_2} - \omega\|$, then $\mathbb{I}(\omega_{N_2} \in \mathcal{S}(k_2, \omega; \{\omega_j\}_{j=1}^{N_2})) = 1$, $\mathbb{I}(\omega^1_{[k_1]} \in \mathcal{S}(k_2, \omega; \{\omega_j\}_{j=1}^{N_2})) = 0$, and $\mathbb{I}(\omega^1_{[i]} \in \mathcal{S}(k_2, \omega; \{\omega_j\}_{j=1}^{N_2}))) = 1$ for $i = 1, 2, ..., k_1 - 1$. By Assumptions A5 and B6,

$$h^{k_2}_{\omega,2}(x) - h^{k_1}_{\omega,1}(x) = \frac{1}{k_2} H(x, \xi_{N_2}) - \frac{1}{k_1} H(x, \xi^1_{[k_1]}) \geq 0 - \frac{M_h}{k_1},\ \forall\ x \in X. \tag{47}$$

Hence, it follows from (47) that $h^{k_2}_{\omega,2}(x) \geq \hat{h}^{k_1}_{\omega,1}(x) - \frac{M_h}{k_1}$, $\forall\ x \in X$, which further implies that

$$h^{k_2}_{\omega,2}(x) \geq \max\{\hat{h}^{k_2}_{\omega,1,\nu+1}(x) - \frac{M_h}{k_1}, \mathcal{C}^1_{\omega,2}(x), \mathcal{C}^2_{\omega,2}(x)\} = \hat{h}^{k_2}_{\omega,2,1}(x)\ \forall\ x \in X.$$

Finally, let us consider the case when $k_2 > k_1$. Let

$$\bar{S}_2 = \{1 \leq i \leq N_2 : \omega_i \notin \mathcal{S}(k_1, \omega; \{\omega_j\}_{j=1}^{N_1}), \omega_i \in \mathcal{S}(k_2, \omega; \{\omega_j\}_{j=1}^{N_2})\}$$

denote the set of indices of covariates so that they are not in the first kNN set, $\mathcal{S}(k_1, \omega; \{\omega_j\}_{j=1}^{N_1})$ but in the second kNN set, $\mathcal{S}(k_2, \omega; \{\omega_j\}_{j=1}^{N_2})$. $N_2 = N_1 + 1$ implies $\mathbb{I}(\omega^1_{[i]} \in \mathcal{S}(k_2, \omega; \{\omega_j\}_{j=1}^{N_2}))) = 1,\ i = 1, 2, ..., k_1$. which further implies that

$$\begin{aligned} h^{k_2}_{\omega,2}(x) &= \frac{1}{k_2} \sum_{i \in \bar{S}_2} H(x, \xi_i) + \frac{1}{k_2} \sum_{i=1}^{k_1} H(x, \xi^1_{[i]}) \\ &\geq 0 + \frac{k_1}{k_2} \frac{1}{k_1} \sum_{i=1}^{k_1} H(x, \xi^1_{[i]}) \geq \frac{k_1}{k_2} \hat{h}^{k_2}_{\omega,1,\nu+1}(x). \end{aligned} \tag{48}$$

Hence, this shows that $h^{k_2}_{\omega,2}(x) \geq \max\{\frac{k_1}{k_2}\hat{h}^{k_2}_{\omega,1,\nu+1}(x), \mathcal{C}^1_{\omega,2}(x), \mathcal{C}^2_{\omega,2}(x)\} = \hat{h}^{k_2}_{\omega,2,1}(x)$, $\forall\, x \in X$.

Case II: $\ell = 1$ and the inner loop is not terminated. By the convexity of $h^{k_1}_{\omega,1}(x)$, we have $h^{k_1}_{\omega,1}(x_1^{\frac{1}{2},2}) + u_{\omega,1}(x_1^{\frac{1}{2},2})^\top(x - x_1^{\frac{1}{2},2}) \leq h^{k_1}_{\omega,1}(x)$, $\forall\, x \in X$. Hence, it is obvious that $\hat{h}^{k_1}_{\omega,1,2}(x) = \max\{\hat{h}^{k_1}_{\omega,1,1}(x), h_{\omega,1}(x_1^{\frac{1}{2},2}) + u_{\omega,1}(x_1^{\frac{1}{2},2})^\top(x - x_1^{\frac{1}{2},2})\} \leq h^{k_1}_{\omega,1}(x)$, $\forall\, x \in X$. Then one can mimic the same procedure to show $\hat{h}^{k_1}_{\omega,1,\nu}(x) \leq h^{k_1}_{\omega,1}(x)$ for all $\nu \geq 1$ until the inner loop is terminated.

Case III: $\ell > 1$ and the inner loop is terminated. By induction, suppose that $\hat{h}^{k_\ell}_{\omega,\ell,\nu+1}(x) \leq h^{k_\ell}_{\omega,\ell}(x)$ after the inner loop. The proof in this case is similar to Case I.

Case IV: $\ell > 1$ and the inner loop is not terminated. The proof is similar to Case II. ■

Proposition 11. *Suppose Assumptions A1 - A6 and B5 - B6 hold.* $\{x^\ell\}_\ell$ *and* $\{\bar{x}^\ell\}_\ell$ *generated by Algorithm 5 satisfies the following relation:*

$$\zeta^{k_\ell}_{\omega,\ell}(x^{\ell+1}) - \zeta^{k_\ell}_{\omega,\ell}(\bar{x}^\ell) \leq -\frac{c}{4}\|x^{\ell+1} - \bar{x}^\ell\|^2. \tag{49}$$

Proof. The proof is similar to Proposition 11. ■

Proposition 12. *Suppose Assumptions A1 - A6 and B5 - B6 hold. Then*

$$\limsup_{\ell\to\infty} \zeta^{k_\ell}_{\omega,\ell}(x^{\ell+1}) - \zeta^{k_\ell}_{\omega,\ell}(\bar{x}^\ell) = 0 \ w.p.1. \tag{50}$$

Proof. The proving strategy is the same as the proof of Proposition 7. ■

Proposition 13. *Suppose Assumptions A1 - A6 hold and B5 - B6 hold. Then with probability one, NSD-MM (Algorithm 5) produces a subsequence,* $\{\bar{x}^\ell\}_{\ell\in\mathcal{L}}$*, such that any accumulation point of* $\{x^\ell\}_{\ell\in\mathcal{L}}$ *(i.e.,* $\mathcal{L}' \subseteq \mathcal{L}$*,* $\lim\limits_{\ell(\in\mathcal{L}')\to\infty} x^\ell = x^\infty$*) satisfies the following relation:*

$$\lim_{\ell(\in\mathcal{L}')\to\infty} x^{\ell+1} = \lim_{\ell(\in\mathcal{L}')\to\infty} x^\ell = x^\infty \in X, \tag{51}$$

and

$$\limsup_{\ell(\in\mathcal{L}')\to\infty} (\hat{h}^{k_\ell}_{\omega,\ell,\nu_\ell})'(x^{\ell+1}; x - x^{\ell+1}) \leq h'_\omega(x^\infty; x - x^\infty), \ \forall\, x \in X. \tag{52}$$

Proof. The proving strategy is similar to the proof of Proposition 4. The only differences are (1) we apply Proposition 12 and Proposition 11 to get that there exists a subsequence, $\{\bar{x}^\ell\}_{\ell\in\mathcal{L}}$, such that $\lim\limits_{\ell(\in\mathcal{L}')\to\infty} \|x^{\ell+1} -$

$\bar{x}^\ell\| = 0$ $w.p.1$ and (2) we utilize pointwise convergence of kNN estimator (Theorem 3) and the equicontinuity of $\{h_\ell^{k_\ell}\}_\ell$ to derive the uniform convergence of $\{h_\ell^{k_\ell}\}_\ell$ on X. ■

Theorem 5. *Suppose Assumptions A1 - A6 and B5 - B6 hold. Further assume either Assumption A7 or A8 holds. Then, with probability one, NSD-MM (Algorithm 5) produces a subsequence, $\{\bar{x}^\ell\}_{\ell\in\mathcal{L}}$, such that any accumulation point of $\{x^\ell\}_{\ell\in\mathcal{L}}$ is a d-stationary point of (45).*

Proof. **Case 1: Assumption About 7 holds.** By Proposition 12 and Proposition 11, there exists a subsequence, $\{\bar{x}^\ell\}_{\ell\in\mathcal{L}}$, such that $\lim_{\ell(\in\mathcal{L}')\to\infty}\|x^{\ell+1}-\bar{x}^\ell\| = 0$ $w.p.1$. Since $\{\bar{x}^\ell\}_{\ell\in\mathcal{L}} \subset X$ and X is compact, the accumulation point of $\{\bar{x}^\ell\}_{\ell\in\mathcal{L}}$ exists, say $\mathcal{L}' \subseteq \mathcal{L}$, $\lim\limits_{\ell(\in\mathcal{L}')\to\infty}\bar{x}^\ell = \lim\limits_{\ell(\in\mathcal{L}')\to\infty} x^{\ell+1} = 0$. By the **Candidate Update** of Algorithm 5, we have the optimality as follows: for any $x \in X$,

$$\begin{aligned}0 \le \frac{1}{k_\ell}\sum_{i=1}^{N_\ell}\mathbb{I}(\omega_i \in \mathcal{S}(k_\ell,\omega;\{\omega_i\}_{i=1}^{N_\ell}))[\nabla_x F(\bar{x}^\ell,\xi_i)^\top(x-x^{\ell+1})]\\ + (\hat{h}^{k_\ell}_{\omega,\ell,\nu_\ell})'(x^{\ell+1};x-x^{\ell-1}) + c(x^{\ell+1}-\bar{x}^\ell)^\top(x-x^{\ell+1}).\end{aligned} \tag{53}$$

Let $\ell(\in\mathcal{L}')\to\infty$, by the pointwise convergence of kNN estimator (Theorem 3) and Theorem 4, we

$$\begin{aligned}&\lim_{\ell(\in\mathcal{L}')\to\infty}\frac{1}{k_\ell}\sum_{i=1}^{N_\ell}\mathbb{I}(\omega_i \in \mathcal{S}(k_\ell,\omega;\{\omega_i\}_{i=1}^{N_\ell}))\nabla F(\bar{x}^\ell,\xi_i)^\top(x-x^\infty)\\ &= \mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[\nabla F(x^\infty,\tilde{\xi})|\tilde{\omega}=\omega]^\top(x-x^\infty)\\ &= \nabla\mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[F(x^\infty,\tilde{\xi})|\tilde{\omega}=\omega]^\top(x-x^\infty)\\ &= \nabla f_\omega(x^\infty)^\top(x-x^\infty).\end{aligned}$$

On the other hand, by Proposition 13, we have

$$\limsup_{\ell(\in\mathcal{L}')\to\infty}(\hat{h}^{k_\ell}_{\omega,\ell,\nu_\ell})'(x^{\ell+1};x-x^{\ell+1}) \le h'_\omega(x^\infty;x-x^\infty),\ \forall\, x\in X.$$

Thus, it follows from (53) that

$$\begin{aligned}0 &\le \liminf_{\ell(\in\mathcal{L}')\to\infty}\frac{1}{k_\ell}\sum_{i=1}^{N_\ell}\mathbb{I}(\omega_i \in \mathcal{S}(k_\ell,\omega;\{\omega_i\}_{i=1}^{N_\ell}))\nabla_x F(\bar{x}^\ell,\xi_i)^\top(x-x^\infty)\\ &+ \liminf_{\ell(\in\mathcal{L}')\to\infty}(\hat{h}^{k_\ell}_{\omega,\ell,\nu_\ell})'(x^{\ell+1};x-x^{\ell+1}) + \liminf_{l(\in\mathcal{L}')\to\infty}c(x^{\ell+1}-\bar{x}^\ell)^\top(x-x^{\ell+1})\\ &\le \nabla f_\omega(x^\infty)^\top(x-x^\infty) + h'_\omega(x^\infty;x-x^\infty),\end{aligned}$$

which shows that x^∞ is a d-stationary point of (45) with probability one.

Case 2: Assumption A8 holds. In this case, the surrogate function of $F(x,\xi)$ near $x' \in X$ is written as: $G(x,\xi;x') = F(x,\xi) + \nabla_x F(x,\xi)^\top(x-x') + \frac{L}{2}\|x-x'\|^2$. The optimality condition of the **Candidate Update** of Algorithm 5 is

$$\begin{aligned}0 \le \frac{1}{k_\ell}\sum_{i=1}^{N_\ell} \mathbb{I}(\omega_i \in \mathcal{S}(k_\ell,\omega;\{\omega_i\}_{i=1}^{N_\ell}))[\nabla_x F(\bar{x}^\ell,\xi_i)^\top(x-x^{\ell+1})] \\ + (\hat{h}^{k_\ell}_{\omega,\ell,\nu_\ell})'(x^{\ell+1};x-x^{\ell-1}) + (c+L)(x^{\ell+1}-\bar{x}^\ell)^\top(x-x^{\ell+1}).\end{aligned} \tag{54}$$

Then the rest of the proof follows the same procedure of the case 1. ■

4 Numerical Experiment

In this section, we apply SD-MM and NSD-MM algorithms to solve a class of two-stage shipment problems with linear or concave first-stage cost. The algorithms are implemented in the C++ environment, and CPLEX v12.8.0, as a base solver, is used to solve convex quadratic programming problems and obtain the second-stage dual multipliers. All the programs are run on the Macbook Pro 2017 with 3.1 GHz Dual-Core Intel Core i5. The following results show that the SD-MM algorithm solves the two-stage stochastic linear programming problems near the optimum and maintains a stable descent in the objective value, and converges in the nonconvex case. As for the NSD-MM solver, the numerical results show that it maintains stable descent in the objective value but has large solution quality oscillation in a more complicated instance, which agrees with the weaker convergence result shown in Theorem 5.

4.1 Two-Stage SLP

The problem formulation of the two-stage stochastic linear programming problem is formulated as $\min_{x\in X} c^\top x + \mathbb{E}_{\tilde{\xi}}[H(x,\tilde{\xi})]$, where x denotes the decision of the amount of production in each factory, and $H(x,\xi)$ denotes the second-stage recourse function. We apply the SD-MM algorithm to solve eight different instances. LandS, LandS2 and PGP2 are power generation problems, BK19 and RETAIL are shipment planning problems, 4NODE and 20TERM are freight fleet scheduling problems, and SSN is a network expansion problem. The cost coefficients and the transportation network of BK19 are due to the two-stage shipment planning example in Bertsimas and Kallus [6], while the data generation process is different. The demands random variables

Instance name	1^{st} stage variables/constraints	2^{nd} stage variables/constraints	Number of random variables	Magnitude of random variables
LandS	4/2	12/7	1	3
LandS2	4/2	12/7	3	64
PGP2	4/2	16/7	3	576
BK19	4/0	52/16	12	∞
RETAIL	7/0	70/22	7	10^{70}
4NODE	52/14	186/74	12	32768
20TERM	63/3	764/124	40	10^{12}
SSN	89/1	706/175	86	10^{70}

Table 1: Problem Complexity of Two-Stage SLPs

follow the truncated normal (i.e, truncating the non-negative realization of the random variable) with means $(5, 6, 7, 8, 5, 6, 7, 8, 5, 6, 7, 8)$, standard deviations all equal to 1, lower bounds all equal 0, and upper bounds equal to $(10, 12, 14, 16, 10, 12, 14, 16, 10, 12, 14, 16)$. The model parameters and data generation process of the other instances can be seen from Sen and Liu [43]. Table 2 summarizes the computational results of the SD-MM algorithm for solving eight Two-Stage SLP instances. Each instance is replicated ten times. The column of "average inner iterations" shows the average total number iterations of the SD-MM algorithm for solving each instance. 4NODE requires more than eight inner loops per outer iteration, which is the highest. The column "Avg(Obj)" summarizes the average solution qualities in terms of the out-of-sample validated costs. To ensure the sufficient size of the out-of-sample validation set, we output the average of the 95% CI half margin in the column "Avg(95%CI half margin)". The column "Std(Obj)" shows the oscillation of the estimated solutions in terms of the standard deviation of out-of-sample validated costs. According to Table 5 from Sen and Liu [43], we conclude that SD-MM can efficiently solve all of the eight instances near optimum.

Instance name	Outer iterations	Average inner iterations	Avg(Obj)	Std(Obj)	Avg(95%CI half margin)	Time elapse (secs)
LandS	200	369.3	382.12	0.01	1.33	6.77
LandS2	200	317.7	226.95	0.32	1.54	7.12
PGP2	200	573.1	447.89	0.66	1.34	9.90
BK19	100	246.1	729.29	1.32	0.77	2.68
RETAIL	500	671.9	154.14	0.16	0.73	45.29
4NODE	200	1741.3	447.07	0.32	0.05	54.95
20TERM	300	2279	254491.10	112.43	156.25	296.19
SSN	1100	2253.9	10.21	0.11	0.12	961.70

Table 2: Computational Results of Two-Stage SLPs

4.2 Two-Stage SP with Concave First-Stage Cost

In the second class of problems, we apply the SD-MM algorithm to solve 2 instances of two-stage SP with concave first-stage cost. The mathematical formulation of the first instance is $\min_{x \in X} \mathbb{E}_{\tilde{\xi}}[\sum_{i=1}^n k_i(\tilde{\xi}) \log(b_i(\tilde{\xi})x_i + 1) + H(x, \tilde{\xi})]$, which is the modification of BK19, which is referred to as BK19(sto klogbx). The second instance is based on the 4NODE, where the second-stage subproblem remains the same and the objective function becomes $\min_{x \in X} \sum_{i=1}^n k_i \log(b_i x_i + 1) + \mathbb{E}_{\tilde{\xi}}[H(x, \tilde{\xi})]$. We shall refer to this instance as 4NODE(det klogbx). For each instance, we replicate the entire process ten times. In both graphs of Figure 2, the solid black line and the shadow area show the average performance and the distribution of the ten replications of the SD-MM algorithm, respectively. In 4NODE(det klogbx), the solution quality stabilizes when the sample size increases to 40. Since the dimension of the dual in the second-stage problem of 4NODE(det klogbx) is 74 while the one of BK19(sto klogbx) is only 16 and the dimension of the first-stage decision is much larger, this explains why the shadow area of 4NODE(det klogbx) does not diminish to 0 while the shadow area of BK19(sto klogbx) can.

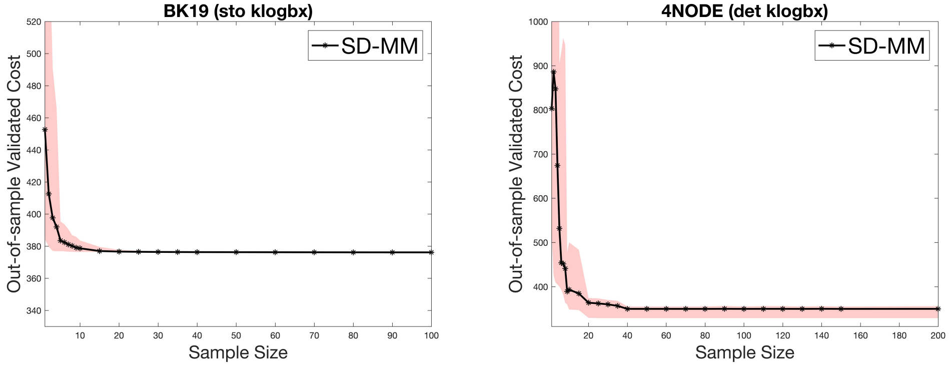

Figure 2: Computational results of SDMM in BK19(sto klogbx) and 4NODE(det klogbx)

4.3 Two-Stage PSP

We reuse the model parameters in PBK19(sto klogbx) and 4NODE(det klogbx) and design new data generation process of the tuple, $(\tilde{\omega}, \tilde{\xi})$ to showcase the computational performance of the NSD-MM algororithm. The mathematical formulation of the first instance, which we refer to as PBK19(sto klogbx), is: $\min_{x \in X} \mathbb{E}_{\tilde{\xi} \sim \mu_{\tilde{\xi}|\omega}}[\sum_{i=1}^n k_i(\tilde{\xi}) \log(b_i(\tilde{\xi})x_i + 1) + H(x, \tilde{\xi}) | \tilde{\omega} = \omega]$. To the end of this section, we let $v_{(i)}$ denote the i^{th} component of the vector v. As for the data generation process, $\omega_(i)$ follows Uniform(0.5,1.5) for $i = 1, 2, 3,$

and the i^{th} component of the demand is:

$$e_i(\xi) = \max\{0, A_i^T(\omega + \delta_i/4) + (B_i^T\omega)\epsilon_i\} \quad i = 1, 2, ..., 12,$$

where δ_i and ϵ_i are white noises, and A_i, B_i are constant vectors. $k(\xi)$ follows $k_{(i)}(\xi) = C_i^\top\omega + \tau_1 + \bar{k}$, where τ_1 is truncated standard normal random variable with (lower bound,upper bound)=(-0.5,0.5) and C_i is a constant vector, for $i = 1, 2, 3, 4$. $b(\xi)$ follows $b(\xi) = \omega_1 + \tau_2 + \bar{b}$, $i = 1, 2, 3, 4$, where τ_1 is a truncated standard normal random variable with (lb,ub) = (-0.1,0.1) and $\bar{b}$ is a constant. We feed the NSD-MM solver with dataset $\{(\omega_i, e_i(\xi), k_i(\xi)), b_i(\xi)\}_{i=1}^{\mathcal{N}}$ and the observation of the predictor $\omega = (1, 1, 1)$. We replicate the data generation and solving processes 10 times and output the average and distribution of the solution qualities evaluated by the out-of-sample validation set in left sub graph of the Figure 3. It shows that the estimated solution of NSD-MM manage to converge in this case.

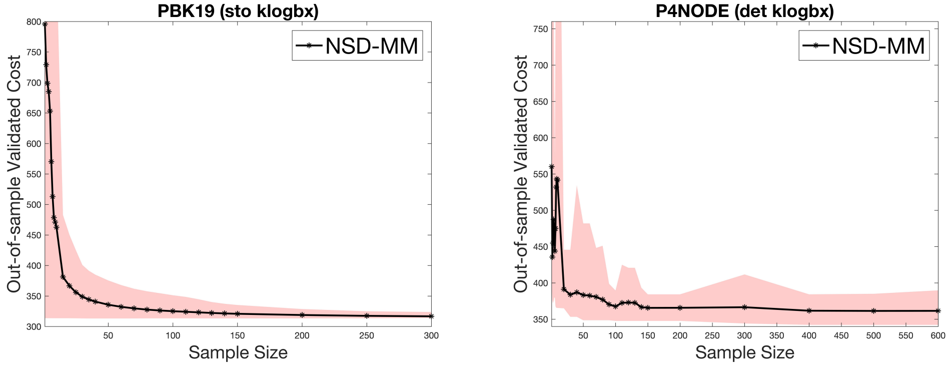

Figure 3: Computational results of NSD-MM in PBK19(sto klogbx) and P4NODE(det klogbx)

The mathematical formulation of the second instance, which we refer to as P4NODE(det klogbx), is $\min\limits_{x\in X} \sum_{i=1}^{n} k_i \log(b_i x_i + 1) + \mathbb{E}_{\tilde{\xi}\sim\mu_{\tilde{\xi}|\omega}}[H(x, \tilde{\xi})|\tilde{\omega} = \omega]$. ω_i follows uniform(0,1), the vector e in the right hand side of the second-stage problem follows the following data generation process:

$$e_i(\xi) = Z_i^\top\omega + \bar{e}_i + \epsilon_i,\ \ i = 1, 2, \ldots, 12,$$

where ϵ_i is truncated standard normal random variable with (lower bound,upper bound)=(-0.5,0.5), Z_i is a constant vector, and $\bar{e}_i$ is a constant scalar. Then we feed the NSD-MM solver with dataset $\{(\omega_i, e_i(\xi))\}_{i=1}^{\mathcal{N}}$ and the observation of the predictor $\omega = (0.4, 0.6)$. Again, we replicate the data generation and solving processes ten times and then plot the average solution qualities of ten replications in the right subgraph of the Figure 3. Its large shadow area in the case of P4NODE(det klogbx) is possibly due to the higher problem complexity and the weak convergence result of NSD-MM as shown in Theorem 5.

5 Conclusions

In this paper, we have designed a fusion of the SD algorithm and MM algorithm to solve a class of non-smooth nonconvex stochastic program. The proposed SD-MM algorithm has a unique property that successively approximates the second-stage cost-to-go function (convex component) from below and approximates the concave or L-smooth first-stage component from above. This technique exploits the historic approximations of the sample-based convex component to estimate the current sample-based convex component. As a result, the SD-MM algorithm can store the memory of the functional estimates in the previous iterations and thus use the memory to attain the next iterate. Moreover, we further equip the SD-MM algorithm with kNN estimation to solve nonconvex predictive stochastic programming problems.

We are aware that the objective function in the problems discussed can also be expressed as an expectation of a smooth concave function and a convex piecewise linear function (i.e., $H(x, \xi) = \max_{1 \leq j \leq J} \{\beta_j(\xi)^\top x + \alpha_j(\xi)\}$), which may open up future applications of the SD-MM algorithm in the machine learning problems. In the future, one interesting direction is to derive the convergence rate of the SD-MM algorithm. Another interesting direction of the research is to study the finite-sample performance of the SD-MM algorithm. The third direction is to merge the methodology of the SD-MM with multistage Benders decomposition to solve multistage stochastic linear program with infinite trials. We hope our methodology provides a new viewpoint on the inexact local approximation of the objective (i.e., neitherThey upper bound nor lower bound) for efficiently looking for d-stationary solutions to the nonconvex non-smooth stochastic programming problems.

Acknowledgments.

This research is supported by the Office of Naval Research Grant N00014-20-1-2077.

References

[1] L. T. H. An, H. Van Ngai, P. D. Tao, and L. H. P. Hau, *Stochastic difference-of-convex algorithms for solving nonconvex optimization problems*, arXiv preprint arXiv:1911.04334, (2019).

[2] G.-Y. Ban and C. Rudin, *The big data newsvendor: Practical insights from machine learning*, Operations Research, 67 (2019), pp. 90–108.

[3] G. Barbarosoğlu and Y. Arda, *A two-stage stochastic programming framework for transportation planning in disaster response*, Journal of the operational research society, 55 (2004), pp. 43–53.

[4] A. Beck, *First-order methods in optimization*, SIAM, 2017.

[5] D. P. Bertsekas, *Incremental proximal methods for large scale convex optimization*, Mathematical programming, 129 (2011), pp. 163–195.

[6] D. Bertsimas and N. Kallus, *From predictive to prescriptive analytics*, Management Science, (2019).

[7] D. Bertsimas and C. McCord, *From predictions to prescriptions in multistage optimization problems*, arXiv preprint arXiv:1904.11637, (2019).

[8] D. C. Cafaro and I. E. Grossmann, *Alternate approximation of concave cost functions for process design and supply chain optimization problems*, Computers & chemical engineering, 60 (2014), pp. 376–380.

[9] E. Çinlar, *Probability and stochastics*, vol. 261, Springer Science & Business Media, New York, 2011.

[10] Y. Cui and J.-S. Pang, *Modern Nonconvex Nondifferentiable Optimization*, SIAM, 2021.

[11] L. Devroye, L. Gyorfi, A. Krzyzak, and G. Lugosi, *On the strong universal consistency of nearest neighbor regression function estimates*, The Annals of Statistics, 22 (1994), pp. 1371–1385.

[12] S. Diao and S. Sen, *Distribution-free algorithms for learning enabled optimization with non-parametric estimation*, Optimization Online, (2020).

[13] R. Durrett, *Probability: theory and examples*, vol. 49, Cambridge university press, 2019.

[14] A. N. Elmachtoub and P. Grigas, *Smart "predict, then optimize"*, Management Science, 68 (2022), pp. 9–26.

[15] E. Fix and J. L. Hodges Jr, *Discriminatory analysis-nonparametric discrimination: consistency properties*, tech. rep., California Univ Berkeley, 1951.

[16] L. Györfi, M. Kohler, A. Krzyzak, H. Walk, et al., *A distribution-free theory of nonparametric regression*, vol. 1, Springer, 2002.

[17] J. L. Higle and S. Sen, *Stochastic decomposition: An algorithm for two-stage linear programs with recourse*, Mathematics of operations research, 16 (1991), pp. 650–669.

[18] ——, *Finite master programs in regularized stochastic decomposition*, Mathematical Programming, 67 (1994), pp. 143–168.

[19] C. Hu, G. Jain, P. Zhang, C. Schmidt, P. Gomadam, and T. Gorka, *Data-driven method based on particle swarm optimization and k-nearest neighbor regression for estimating capacity of lithium-ion battery*, Applied Energy, 129 (2014), pp. 49–55.

[20] Y. Hu, N. Kallus, and X. Mao, *Fast rates for contextual linear optimization*, arXiv preprint arXiv:2011.03030, (2020).

[21] G. Huang and D. P. Loucks, *An inexact two-stage stochastic programming model for water resources management under uncertainty*, Civil Engineering Systems, 17 (2000), pp. 95–118.

[22] R. Kannan, G. Bayraksan, and J. Luedtke, *Heteroscedasticity-aware residuals-based contextual stochastic optimization*, arXiv preprint arXiv:2101.03139, (2021).

[23] R. Kannan, G. Bayraksan, and J. R. Luedtke, *Data-driven sample average approximation with covariate information*, Optimization Online. URL: http://www. optimization-online. org/DB HTML/2020/07/7932. html, (2020).

[24] H. Konno and A. Wijayanayake, *Portfolio optimization problem under concave transaction costs and minimal transaction unit constraints*, Mathematical Programming, 89 (2001), pp. 233–250.

[25] S. C. Kumbhakar and C. F. Parmeter, *Estimation of hedonic price functions with incomplete information*, Empirical Economics, 39 (2010), pp. 1–25.

[26] H. A. Le Thi and T. Pham Dinh, *Dc programming and dca: thirty years of developments*, Mathematical Programming, 169 (2018), pp. 5–68.

[27] H. Li and Y. Cui, *A decomposition algorithm for two-stage stochastic programs with nonconvex recourse*, arXiv preprint arXiv:2204.01269, (2022).

[28] J. Liu, Y. Cui, and J.-S. Pang, *Solving nonsmooth nonconvex compound stochastic programs with applications to risk measure minimization*, arXiv preprint arXiv:2004.14342, (2020).

[29] J. Liu, Y. Cui, J.-S. Pang, and S. Sen, *Two-stage stochastic programming with linearly bi-parameterized quadratic recourse*, SIAM Journal on Optimization, 30 (2020), pp. 2530–2558.

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

[31] J. Mairal, *Stochastic majorization-minimization algorithms for large-scale optimization*, Advances in Neural Information Processing Systems, 26 (2013).

[32] ——, *Incremental majorization-minimization optimization with application to large-scale machine learning*, SIAM Journal on Optimization, 25 (2015), pp. 829–855.

[33] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, *Online learning for matrix factorization and sparse coding.*, Journal of Machine Learning Research, 11 (2010).

[34] E. Mangalova and E. Agafonov, *Wind power forecasting using the k-nearest neighbors algorithm*, International Journal of Forecasting, 30 (2014), pp. 402–406.

[35] C. Martins-Filho and O. Bin, *Estimation of hedonic price functions via additive nonparametric regression*, Empirical economics, 30 (2005), pp. 93–114.

[36] T. Oladunni and S. Sharma, *Hedonic housing theory—a machine learning investigation*, in 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2016, pp. 522–527.

[37] A. Owusu-Ansah, *A review of hedonic pricing models in housing research*, Journal of International Real Estate and Construction Studies, 1 (2011), p. 19.

[38] J.-S. Pang, M. Razaviyayn, and A. Alvarado, *Computing b-stationary points of nonsmooth dc programs*, Mathematics of Operations Research, 42 (2017), pp. 95–118.

[39] D. Phan and S. Ghosh, *Two-stage stochastic optimization for optimal power flow under renewable generation uncertainty*, ACM Transactions on Modeling and Computer Simulation (TOMACS), 24 (2014), pp. 1–22.

[40] A. B. Philpott and Z. Guan, *On the convergence of stochastic dual dynamic programming and related methods*, Operations Research Letters, 36 (2008), pp. 450–455.

[41] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, *A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks*, Mathematical Programming, 157 (2016), pp. 515–545.

[42] S. Sen, R. D. Doverspike, and S. Cosares, *Network planning with random demand*, Telecommunication systems, 3 (1994), pp. 11–30.

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

[44] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on stochastic programming: modeling and theory*, SIAM, 2021.

[45] B. L. Smith and M. J. Demetsky, *Short-term traffic flow prediction models-a comparison of neural network and nonparametric regression approaches*, in Proceedings of IEEE international conference on systems, man and cybernetics, vol. 2, IEEE, 1994, pp. 1706–1709.

[46] C. J. Stone, *Consistent nonparametric regression*, The annals of statistics, (1977), pp. 595–620.

[47] P. D. Tao and L. T. H. An, *Convex analysis approach to dc programming: theory, algorithms and applications*, Acta mathematica vietnamica, 22 (1997), pp. 289–355.

[48] H. Walk, *A universal strong law of large numbers for conditional expectations via nearest neighbors*, Journal of Multivariate Analysis, 99 (2008), pp. 1035–1050.

[49] A. D. Witte, H. J. Sumka, and H. Erekson, *An estimate of a structural hedonic price model of the housing market: an application of rosen's theory of implicit markets*, Econometrica: Journal of the Econometric Society, (1979), pp. 1151–1173.