On Data-Driven Prescriptive Analytics with Side Information: A Regularized Nadaraya-Watson Approach

Chin Pang Ho* Grani A. Hanasusanto†

Abstract

We consider generic stochastic optimization problems in the presence of side information which enables a more insightful decision. The side information constitutes observable exogenous covariates that alter the conditional probability distribution of the random problem parameters. A decision maker who adapts her decisions according to the observed side information solves an optimization problem where the objective function is specified by the conditional expectation of the random cost. If the joint probability distribution is unknown then the conditional expectation can be approximated in a data-driven manner using the Nadaraya-Watson (NW) kernel regression. While the emerging approximation scheme has found successful applications in diverse decision problems under uncertainty, it is largely unknown whether the scheme can provide any reasonable out-of-sample performance guarantees. In this paper, we establish guarantees for the generic problems by leveraging techniques from moderate deviations theory. The new theoretical result motivates the design of an effective regularization scheme via empirical conditional standard deviation. We highlight the performance of the regularized NW approximation through an example in portfolio management.

Keywords: stochastic optimization; side information; Nadaraya-Watson estimator; moderate deviation principles; large deviation principles

1 Introduction

In the presence of uncertainty, decisions can often be improved by taking into account the side information, such as the weather condition, interest rate, exchange rates, past prices and demands, volatility indices, etc., that provides a more precise description about the uncertain problem parameters. In the stochastic optimization setting, the side information corresponds to observable exogenous covariates $(\gamma_1, \ldots, \gamma_p)$ that may reshape the conditional probability distribution of the random problem parameters $(\tilde{\xi}_1, \ldots, \tilde{\xi}_q)$. A decision maker who is prescribed with full knowledge about the joint distribution of the random vectors $\tilde{\boldsymbol{\gamma}} := (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p)$ and $\tilde{\boldsymbol{\xi}} := (\tilde{\xi}_1, \ldots, \tilde{\xi}_q)$ endeavors to solve the stochastic optimization problem

$$\min_{\boldsymbol{x} \in \mathcal{X}} \ \mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] := \mathbb{E}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}}) \,|\, \tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}]. \tag{$\mathcal{S}$}$$

*Imperial College Business School, Imperial College London, London, SW7 2AZ, UK. Email: c.ho12@imperial.ac.uk.

†Graduate Program in Operations Research and Industrial Engineering, The University of Texas at Austin, Austin, TX, 78712-1591, USA. Email: grani.hanasusanto@utexas.edu.

Here, the vector $\boldsymbol{x} \in \mathbb{R}^d$ comprises all decision variables and the objective function is specified by the conditional expectation of the random cost $\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})$ given the side information $\boldsymbol{\gamma}$.

In most situations of practical interest, the joint distribution of $(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\xi}})$ is unknown and only past historical data $\{(\boldsymbol{\gamma}^1, \boldsymbol{\xi}^1), \ldots, (\boldsymbol{\gamma}^n, \boldsymbol{\xi}^n)\}$ is available to infer the conditional distribution of $\tilde{\boldsymbol{\xi}}$, and to estimate the conditional expectation in $(\mathcal{S})$. A popular approximation scheme is the Nadaraya-Watson (NW) kernel regression [Nad64, Wat64]

$$\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] = \frac{\sum_{i=1}^n \mathcal{K}_h(\boldsymbol{\gamma} - \boldsymbol{\gamma}^i)\ell(\boldsymbol{x}, \boldsymbol{\xi}^i)}{\sum_{i=1}^n \mathcal{K}_h(\boldsymbol{\gamma} - \boldsymbol{\gamma}^i)}, \tag{1}$$

where the kernel function $\mathcal{K}_h$ is defined through the multivariate Gaussian density function

$$\mathcal{K}_h(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{\|\boldsymbol{\theta}\|_2^2}{2h^2}\right). \tag{2}$$

This setting encapsulates a popular model in data-driven analytics. Indeed, an extremely large value of the bandwidth parameter h means that the approximation (1) reduces to the unconditional *sample-average approximation* $\frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{x}, \boldsymbol{\xi}^i)$. On the other hand, a very small bandwidth implies that most of the probability mass is assigned to the sample point closest to $\boldsymbol{\gamma}$. The choice $h = O(1/n^{1/(p+4)})$ provides the best balance between bias and variance that yields the minimum expected error [GKKW06].

Using the NW estimator (1), we arrive at the following approximation to the stochastic optimization problem $(\mathcal{S})$:

$$\min_{\boldsymbol{x} \in \mathcal{X}} \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]. \tag{$\mathcal{NW}$}$$

This approximation is first developed by Hannah et al. [HPB10]. Bertsimas and Kallus [BK14] provide a generalization by considering different ways of constructing the empirical conditional expectation. They further establish that the resulting approximations are asymptotically consistent, meaning that the approximations converge to the true conditional expectation as the sample size grows. Solutions to the problem $(\mathcal{NW})$, however, exhibit an optimistic bias if the sample size is small. To mitigate this overfitting effect, Hanasusanto and Kuhn [HK13] propose a robust version that minimizes a worst-case empirical conditional expectation in view the most adverse weight vector close to the nominal one generated by the NW estimator. Bertsimas and Van Parys [BVP17] propose an alternative robust scheme whose solutions enjoy a limited disappointment on the *bootstrap data*.

Despite the practical significance of stochastic optimization problem $(\mathcal{S})$, there is so far an incomplete picture about the properties of the state-of-the-art approximation schemes. Although the NW approximation $(\mathcal{NW})$ is known to be asymptotically consistent [BK14], it is unknown whether the scheme could provide out-of-sample performance guarantees for solutions to the generic problems. An alternative scheme which optimizes over parametric *decision rules*, such as linear or quadratic functions in $\boldsymbol{\gamma}$, can generate finite-sample performance bounds [BK14, BR18, BMD18]. However, the scheme is less attractive because it is not asymptotically consistent, meaning that we

cannot produce results that would parallel those of sample-average approximation in the classical setting of stochastic optimization without side information [KSH02, SDR09]. In [BR18], the authors apply both the NW and decision rule approximations to the single-item newsvendor problem, and derive finite-sample performance guarantees for the respective solutions. Unfortunately, the bound for the NW approximation relies inconveniently on an optimal solution to the corresponding linear decision rule problem. An alternative bound derived in [BK14] holds only for the bootstrap data which is generated via resampling from the empirical distribution. Although encouraging, the bound might be deceiving because it does not depend on the dimension p of the exogenous covariate vector $\tilde{\boldsymbol{\gamma}}$.

This paper takes a first step toward gaining a complete understanding about the out-of-sample performance of approximation ($\mathcal{NW}$). By leveraging techniques from *large and moderate deviations theory*, we derive for the first time out-of-sample performance guarantees of the generic problem. Our result indicates that the out-of-sample errors of the approximation scale with $O(\sqrt{1/(nh^p)})$. In contrast to the result in [BR18] for a single-item newsvendor problem, our guarantees hold independently of optimal solutions to the corresponding linear decision rule problems and conform with the best bandwidth parameter scaling $h = O(1/n^{1/(p+4)})$ suggested in the literature. As a byproduct of our new theoretical result, we identify a suitable regularization term in *empirical conditional standard deviation*. If this term is small then our guarantees imply that the out-of-sample errors are of the smaller rate $\sim O(1/(nh^p))$. Thus, the regularization term will encourage an optimal solution that yields low generalization errors. Empirical results in the context of portfolio optimization demonstrate the superiority of our new regularized NW approximation over the state-of-the-art linear decision rule scheme.

In this paper, we assume the following regularity conditions:

(**A1**) The support Ξ of the random vector $\tilde{\boldsymbol{\xi}}$ is compact, while the loss function $\ell(\boldsymbol{x}, \boldsymbol{\xi})$ is measurable and takes value in the interval $[0, 1]$ for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\xi} \in \Xi$.

(**A2**) The density function $f(\boldsymbol{\gamma}, \boldsymbol{\xi})$ is twice differentiable with continuous and bounded partial derivatives.

(**A3**) The bandwidth parameter h for the kernel function $\mathcal{K}_h$ is scaled such that $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} nh_n^p = \infty$.

The assumptions about the support set and the loss function in (**A1**) are typical in the literature. Here, we do not impose any restriction on the size and structure of the support set other than its compactness. If the loss function is bounded, then one can simply apply scaling and translation so that it takes value in the interval $[0, 1]$. The assumptions about the density function in (**A2**) are standard regularity conditions in kernel density and kernel regression estimations. They ensure that the conditional distribution of $\tilde{\boldsymbol{\xi}}$ given the side information $\boldsymbol{\gamma}$ can be inferred reasonably well using the historical observations. The assumption about the bandwidth parameter h in (**A3**) ensures that the estimator (1) is asymptotically consistent [GKKW06, Sil86].

Large and moderate deviations theory

Large deviations theory studies the tail behavior of sequences of random variables. It characterizes the exponential decay rate of the probability that a random variable in the sequence realizes on any particular rare event. Formally, we say that the sequence of random variables $\{\tilde{z}_n\}_{n\in\mathbb{N}}$ satisfies a large deviation principle with speed ν_n and rate function $I : \mathbb{R} \to [0, +\infty]$ if

$$\liminf_{n\to\infty} \frac{1}{\nu_n} \log \mathbb{P}\left(\tilde{z}_n \in \mathcal{O}\right) \geq - \inf_{y\in\mathcal{O}} I(y) \quad \text{and} \quad \limsup_{n\to\infty} \frac{1}{\nu_n} \log \mathbb{P}\left(\tilde{z}_n \in \mathcal{C}\right) \leq - \inf_{y\in\mathcal{C}} I(y), \tag{3}$$

for every open subset $\mathcal{O}$ and closed subset $\mathcal{C}$ of $\mathbb{R}$, respectively. If the random variable is defined as the average $\tilde{z}_n = \frac{1}{n}\sum_{i=1}^n \tilde{r}_i$ of i.i.d. random variables $\tilde{r}_i$, $i \in \mathbb{N}$, with a finite logarithmic moment generating function $\Lambda(t) = \mathbb{E}[\exp(t\tilde{r}_1)] < +\infty$, then we obtain the Cramer's theorem which states that the sequence $\{\tilde{z}_n\}_{n\in\mathbb{N}}$ obeys a large deviation principle with speed $\nu_n = n$ and rate $I(y) = \sup_{t\geq 0}(ty - \Lambda(t))$. The inequalities in (3) thus imply that for large enough n the probability that $\tilde{z}_n$ takes value within the rare event set $\{z : z \geq y\}$, with $y > \mathbb{E}[\tilde{r}_1]$, is roughly equal to $\exp(-nI(y))$. That is, it decays exponentially fast in n at the rate $I(y)$. Note that the rate function depends on the particular distribution of the random variable $\tilde{r}_1$. From the central limit theorem, however, we know that the distribution of the renormalized average $\sqrt{n}\tilde{z}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{r}_i$ is asymptotically normal, which admits a succinct description through the first and second-order moments of $\tilde{r}_1$.

Moderate deviations theory delineates the intermediate cases between the two extremes of large deviations theory and central limit theorem. The theory studies situations where the sequence $\{a_n\tilde{z}_n\}_{n\in\mathbb{N}}$ obeys a large deviation principle with the *same* rate function for a certain range of renormalization parameters $a_n \to \infty$. The theory often provides a result that combines both large deviations theory and central limit theorem. Analogous to the central limit behavior, the rate function in a moderate deviation principle is typically *analytical*, requiring only limited information about the distribution, such as the variance. However, we also observe an exponential decay rate characteristic of results in large deviations theory. In the case of i.i.d. random variables, we find that if $a_n^2/n \to 0$ as $n \to \infty$ then the sequence $\{a_n\tilde{z}_n\}_{n\in\mathbb{N}}$ obeys a large deviation principle with speed n/a_n^2 and analytical rate function $I(y) = \frac{1}{2}y^2/\sigma^2$, where σ^2 is the variance of the random variable $\tilde{r}_1$ [DZ98, Theorem 3.7.1]. We refer the reader to the references [DZ98, EL03] for a more detailed account on large and moderate deviations theory.

Notation and terminology We use bold letters for vectors, while scalars are printed in regular font. We denote by $\mathbf{e}$ the vector of all ones. Random variables are designated by tilde signs (e.g., $\tilde{\boldsymbol{\xi}}$), while their realizations are represented by the same symbols without tildes (e.g., $\boldsymbol{\xi}$). For any $n \in \mathbb{N}$, we define $[n]$ as the index set $\{1, \ldots, n\}$. We use $\theta(1)$ as a shorthand for the class of functions that have the asymptotic growth rate of $1 + o(1)$. Formally, we say that the function $f : \mathbb{N} \to \mathbb{R}$ satisfies the inclusion $f(n) \in \theta(1)$ if for every $\tau > 0$ there exists $n_\tau \in \mathbb{N}$ such that for all $n \geq n_\tau$ we have $1 - \tau < f(n) < 1 + \tau$. That is, $\lim_{n\to\infty} f(n) = 1$. We define by $\mathcal{SOC}(n + 1) \subseteq \mathbb{R}^{n+1}$ the standard second-order cone: $\boldsymbol{v} \in \mathcal{SOC}(n + 1) \iff \|(v_1, \ldots, v_n)^\top\| \leq v_{n+1}$.

2 Generalization bounds via moderate deviation principles

In this section, we first derive generalization bounds on the NW approximation ($\mathcal{NW}$) for a fixed decision $\boldsymbol{x}$. The result leverages the following moderate deviations theory of the NW estimator by Mokkadem et al. [MPT08, Theorem 2] in the setting of Gaussian kernel functions.

Theorem 1. *[Moderate Deviation Principles] Let the density function $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ satisfy assumption (**A2**). Consider a function $L : \mathbb{R}^q \to \mathbb{R}$ that satisfies the following conditions:*

1. *The function $\boldsymbol{\gamma} \to \int_{\mathbb{R}^q} L(\boldsymbol{\xi})^2 f(\boldsymbol{\gamma}, \boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}$ is continuous at $\boldsymbol{\gamma}$.*
2. *For every $u \in \mathbb{R}$, the function $\boldsymbol{\gamma} \to \int_{\mathbb{R}^q} \exp(uL(\boldsymbol{\xi})) f(\boldsymbol{\gamma}, \boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}$ is bounded and continuous at $\boldsymbol{\gamma}$.*
3. *The function $\boldsymbol{\gamma} \to \int_{\mathbb{R}^q} L(\boldsymbol{\xi}) f(\boldsymbol{\gamma}, \boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}$ is twice differentiable, with continuous and bounded partial derivatives at $\boldsymbol{\gamma}$.*

Then, for any positive sequence $\{a_n\}_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} a_n = \infty, \quad \lim_{n\to\infty} \frac{a_n^2}{nh_n^p} = 0 \quad \text{and} \quad \lim_{n\to\infty} a_n h_n^2 = 0,$$

the sequence $\{a_n(\mathbb{E}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})] - \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})])\}_{n\in\mathbb{N}}$ satisfies a large deviation principle with speed $\nu_n = nh_n^p/a_n^2$ and rate function

$$I_{\boldsymbol{\gamma}}(y) = \frac{y^2 g(\boldsymbol{\gamma})}{\mathbb{V}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})]}, \tag{4}$$

where $g(\boldsymbol{\gamma}) = f(\boldsymbol{\gamma})\pi^{\frac{p}{2}} 2^{\frac{3}{2}p-1}$ is a scaled marginal density of $\tilde{\boldsymbol{\gamma}}$ and $\mathbb{V}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})] = \mathbb{V}[L(\tilde{\boldsymbol{\xi}})|\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}]$ is the conditional variance of $L(\tilde{\boldsymbol{\xi}})$ given the side information $\boldsymbol{\gamma}$. That is, we have

$$\begin{aligned} &\liminf_{n\to\infty} \frac{1}{\nu_n} \log \mathbb{P}\left(a_n\left(\mathbb{E}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})] - \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})]\right) \in \mathcal{O}\right) \geq -\inf_{y\in\mathcal{O}} I_{\boldsymbol{\gamma}}(y) \quad \text{and} \\ &\limsup_{n\to\infty} \frac{1}{\nu_n} \log \mathbb{P}\left(a_n\left(\mathbb{E}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})] - \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})]\right) \in \mathcal{C}\right) \leq -\inf_{y\in\mathcal{C}} I_{\boldsymbol{\gamma}}(y), \end{aligned} \tag{5}$$

for every open subset $\mathcal{O}$ and closed subset $\mathcal{C}$ of $\mathbb{R}$, respectively.

Proof. Suppose assumption (**A2**) and all conditions in the theorem are satisfied. If additionally the kernel function $\mathcal{K}_h(\boldsymbol{\theta})$ satisfies the conditions

$$\int_{\mathbb{R}^p} \boldsymbol{\theta}\mathcal{K}_h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} = \mathbf{0} \quad \text{and} \quad \int_{\mathbb{R}^p} |\theta_j|^\ell \mathcal{K}_h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} < \infty \qquad \forall \ell \in [2]\ \forall j \in [p], \tag{6}$$

then Theorem 2 in [MPT08] implies that the sequence $\{a_n(\mathbb{E}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})] - \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})])\}_{n\in\mathbb{N}}$ obeys a large deviation principle with speed nh_n^p/a_n^2 and rate function

$$I_{\boldsymbol{\gamma}}(y) = \frac{y^2 f(\boldsymbol{\gamma})}{2\mathbb{V}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})] \int_{\mathbb{R}^p} \mathcal{K}_h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}. \tag{7}$$

The Gaussian kernel function (2) satisfies all conditions in (6). The result then follows by substituting the definition of the kernel function (2) for $\mathcal{K}_h(\boldsymbol{\theta})$, and evaluating the integral $\int_{\mathbb{R}^p} \mathcal{K}_h(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$ in (7) analytically. □

Using this theorem, we arrive at our first main result.

Theorem 2. *For any fixed $\boldsymbol{x} \in \mathcal{X}$, we have*

$$\mathbb{P}\left(\left|\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] - \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]\right| \geq \epsilon\right) = \exp\left(-nh_n^p \frac{\epsilon^2 g(\boldsymbol{\gamma})\theta(1)}{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}\right). \tag{8}$$

Proof. We set the function in (5) to $L(\boldsymbol{\xi}) = \ell(\boldsymbol{x},\boldsymbol{\xi})$ and verify that the conditions in Theorem 1 are satisfied. To establish continuity of $\int_\Xi \ell(\boldsymbol{x},\boldsymbol{\xi})^2 f(\boldsymbol{\gamma},\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}$ at $\boldsymbol{\gamma}$, we fix $\epsilon > 0$ and show that there exists $\delta > 0$ such that

$$\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\| \leq \delta \Longrightarrow \left|\int_\Xi \ell(\boldsymbol{x},\boldsymbol{\xi})^2 f(\boldsymbol{\gamma},\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} - \int_\Xi \ell(\boldsymbol{x},\boldsymbol{\xi})^2 f(\boldsymbol{\gamma}',\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}\right| \leq \epsilon. \tag{9}$$

Let $\mu(\Xi)$ be the Lebesgue measure of the support set Ξ. By assumption (**A1**), the following chain of inequalities hold:

$$\begin{aligned}\left|\int_\Xi \ell(\boldsymbol{x},\boldsymbol{\xi})^2 f(\boldsymbol{\gamma},\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} - \int_\Xi \ell(\boldsymbol{x},\boldsymbol{\xi})^2 f(\boldsymbol{\gamma}',\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}\right| &\leq \sup_{\boldsymbol{\xi}\in\Xi} \left|\ell(\boldsymbol{x},\boldsymbol{\xi})^2\right| \sup_{\boldsymbol{\xi}\in\Xi} |f(\boldsymbol{\gamma},\boldsymbol{\xi}) - f(\boldsymbol{\gamma}',\boldsymbol{\xi})|\mu(\Xi) \\ &\leq \sup_{\boldsymbol{\xi}\in\Xi} |f(\boldsymbol{\gamma},\boldsymbol{\xi}) - f(\boldsymbol{\gamma}',\boldsymbol{\xi})|\mu(\Xi).\end{aligned}$$

We now show that there exists $\delta > 0$ such that

$$\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\| \leq \delta \implies \sup_{\boldsymbol{\xi}\in\Xi} |f(\boldsymbol{\gamma},\boldsymbol{\xi}) - f(\boldsymbol{\gamma}',\boldsymbol{\xi})| \leq \epsilon/\mu(\Xi), \tag{10}$$

which proves the claim. Suppose for the sake of contradiction the implication (10) does not hold. That is, for any $\delta > 0$, there exist $\boldsymbol{\gamma}'_\delta$ with $\|\boldsymbol{\gamma} - \boldsymbol{\gamma}'_\delta\| \leq \delta$ and $\boldsymbol{\xi}_\delta \in \Xi$ such that $|f(\boldsymbol{\gamma},\boldsymbol{\xi}_\delta) - f(\boldsymbol{\gamma}'_\delta,\boldsymbol{\xi}_\delta)| > \epsilon/\mu(\Xi)$. By construction, we have $\lim_{\delta\to 0} \boldsymbol{\gamma}'_\delta = \boldsymbol{\gamma}$. Let $\boldsymbol{\xi}^\star$ be a limit point of the sequence $\{\boldsymbol{\xi}_\delta\}$ as $\delta \to 0$. By the compactness of the support set in assumption (**A1**) we have $\boldsymbol{\xi}^\star \in \Xi$. The continuity of the density function in assumption (**A2**) then implies that

$$\epsilon/\mu(\Xi) \leq \lim_{\delta\to 0} |f(\boldsymbol{\gamma},\boldsymbol{\xi}_\delta) - f(\boldsymbol{\gamma}'_\delta,\boldsymbol{\xi}_\delta)| = |f(\boldsymbol{\gamma},\boldsymbol{\xi}^\star) - f(\boldsymbol{\gamma},\boldsymbol{\xi}^\star)| = 0,$$

which is a contradiction because $\epsilon/\mu(\Xi) > 0$. We may thus conclude that the first condition in Theorem 1 is indeed satisfied.

By following the same argument, one can show that $\int_{\mathbb{R}^q} \exp(u\ell(\boldsymbol{x},\boldsymbol{\xi}))f(\boldsymbol{\gamma},\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi}$ is continuous at $\boldsymbol{\gamma}$. The boundedness of the expression holds because $\sup_{\boldsymbol{\xi}\in\Xi} \exp(u\ell(\boldsymbol{x},\boldsymbol{\xi})) \leq \exp(u)$ for every $u \in \mathbb{R}$. Thus, the second condition in Theorem 1 is also satisfied. Finally, by the Leibniz's rule we

have

$$\frac{\partial}{\partial\gamma_i}\int_\Xi \ell(\boldsymbol{x},\boldsymbol{\xi})f(\boldsymbol{\gamma},\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \int_\Xi \ell(\boldsymbol{x},\boldsymbol{\xi})\frac{\partial}{\partial\gamma_i}f(\boldsymbol{\gamma},\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \qquad \forall i\in[p] \qquad \text{and}$$
$$\frac{\partial^2}{\partial\gamma_i\partial\gamma_j}\int_\Xi \ell(\boldsymbol{x},\boldsymbol{\xi})f(\boldsymbol{\gamma},\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} = \int_\Xi \ell(\boldsymbol{x},\boldsymbol{\xi})\frac{\partial^2}{\partial\gamma_i\partial\gamma_j}f(\boldsymbol{\gamma},\boldsymbol{\xi})\mathrm{d}\boldsymbol{\xi} \quad \forall i,j\in[p],$$

where the interchange of the differentiation and the integration operators is valid by the dominated convergence theorem. Thus, in view of our assumption that $\partial f(\boldsymbol{\gamma},\boldsymbol{\xi})/\partial\gamma_i$ and $\partial^2 f(\boldsymbol{\gamma},\boldsymbol{\xi})/(\partial\gamma_i\partial\gamma_j)$ are continuous and bounded, we may apply the same argument to conclude that the third condition in Theorem 1 is also satisfied.

Next, let the closed and the open sets in (5) be defined as $\mathcal{C} = (-\infty,-\epsilon]\cup[\epsilon,\infty)$ and $\mathcal{O} = (-\infty,-\epsilon)\cup(\epsilon,\infty)$, respectively. The function $I_{\boldsymbol{\gamma}}(y)$ is a convex quadratic function centered at 0, which implies that $\inf_{y\in\mathcal{C}} I_{\boldsymbol{\gamma}}(y) = \inf_{y\in\mathcal{O}} I_{\boldsymbol{\gamma}}(y) = I_{\boldsymbol{\gamma}}(\epsilon)$. Thus, we obtain

$$\begin{aligned} -I_{\boldsymbol{\gamma}}(\epsilon) &\le \liminf_{n\to\infty}\frac{1}{\nu_n}\log\mathbb{P}\left(a_n\left(\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]-\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]\right)\in\mathcal{O}\right)\\ &\le \limsup_{n\to\infty}\frac{1}{\nu_n}\log\mathbb{P}\left(a_n\left(\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]-\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]\right)\in\mathcal{C}\right) \le -I_{\boldsymbol{\gamma}}(\epsilon), \end{aligned} \tag{11}$$

which gives rise to the stronger result

$$\frac{1}{\nu_n}\log\mathbb{P}\left(\left|a_n\left(\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]-\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]\right)\right|\ge\epsilon\right) = -I_{\boldsymbol{\gamma}}(\epsilon)+o(1). \tag{12}$$

Multiplying both sides of the inequality with ν_n, taking exponential, and substituting the definition of $G_{\boldsymbol{\gamma}}(\epsilon)$ yield

$$\mathbb{P}\left(\left|a_n\left(\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]-\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]\right)\right|\ge\epsilon\right) = \exp\left(-\frac{\epsilon^2\nu_n g(\boldsymbol{\gamma})}{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}+o(\nu_n)\right).$$

Next, by performing the substitution $\epsilon\leftarrow\epsilon a_n$ and setting $\nu_n = nh_n^p/a_n^2$, we obtain

$$\mathbb{P}\left(\left|\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]-\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]\right|\ge\epsilon\right) = \exp\left(-nh_n^p\frac{\epsilon^2 g(\boldsymbol{\gamma})[1+o(1/a_n^2)]}{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}\right).$$

Since $\lim_{a_n\to\infty} a_n^2 = \infty$, we complete the proof. □

Theorem 2 asserts that, as the sample size grows, the probability that the NW approximation deviates by at least ϵ from the true conditional expectation decays exponentially fast in nh_n^p. Setting the right-hand side of (8) to δ, we arrive at the following result about the out-of-sample errors.

Corollary 1 (Generalization Bound)**.** *For any fixed $\boldsymbol{x}\in\mathcal{X}$, we have*

$$\left|\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]-\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]\right| \le \sqrt{\frac{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}\log\left(\frac{1}{\delta}\right)} = O\left(\sqrt{\frac{1}{nh_n^p}}\right), \tag{13}$$

with probability at least $1-\delta$.

The bound in (13) degrades if the scaled density $g(\boldsymbol{\gamma})$ is small or if the conditional variance $\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$ is large. In the limit where $g(\boldsymbol{\gamma}) \downarrow 0$, there are few historical samples that are close to the given side information, implying that the NW estimator constitutes a poor approximation of the true conditional expectation. On the other hand, a smaller variance indicates that a few data points are sufficient to accurately describe the conditional distribution of $\tilde{\boldsymbol{\xi}}$ given $\boldsymbol{\gamma}$.

Using the best bandwidth parameter scaling $h_n = O(1/n^{1/(p+4)})$ for the multivariate NW estimator [GKKW06, Chapter 5.2], we find that the error bound in (13) diminishes at the rate of $O(1/n^{2/(p+4)})$. Note that we have a dependence on the dimension p, which suggests that the estimator suffers from the *curse of dimensionality*. It may appear here that the dependence on d can be avoided by using the bandwidth scaling $h_n = O(1/n^{1/(cp)})$, where c is a constant greater than 1. In this case, the error decreases at the rate $O(1/n^{(c-1)/(2c)})$ independently of the dimension p. However, as the bandwidth is suboptimal the magnitude of the term $o(1)$ in $\theta(1)$ inflates, implying that the approximation becomes poor when the sample size is small. The result in Theorem 2 indicates that as the sample size grows one may employ the bandwidth scaling $h_n = O(1/n^{1/(cp)})$ to eliminate the dependence on the dimension p.

3 A regularization scheme and its suboptimality bounds

The bound in (13) implies that the out-of-sample errors are negligible if the conditional standard deviation $\sqrt{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}$ is small. This sugggests that a regularization scheme involving the term would ensure a solution with strong generalization bound. As we do not have access to the true conditional variance, we propose to utilize the *empirical* conditional variance as a surrogate:

$$\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] := \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}}) - \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})])^2] = \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})^2] - \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]^2. \tag{14}$$

This setting gives rise to the regularized NW approximation

$$\min_{\boldsymbol{x}\in\mathcal{X}} \ \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})] + \lambda\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}, \tag{$\mathcal{RNW}$}$$

where $\lambda \geq 0$ is a tuning parameter that controls the degree of regularization.

In this section, we aim to establish the properties of the optimal solutions to problem ($\mathcal{RNW}$). We first show that replacing the true conditional variance with its empirical estimate (14) does not significantly weaken the generalization bound derived in Section 2. To this end, we rely on the following lemma.

Lemma 1. *For any fixed* $\boldsymbol{x} \in \mathcal{X}$ *and* $t \in [0,1]$*, we have*

$$\left|\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]} - \sqrt{\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}\right| \leq \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}, \tag{15}$$

with probability at least $1-\delta$.

Proof. By applying Theorem 2 to the input function $(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2$, which also satisfies all conditions in the theorem, we obtain that with a probability at least $1-\delta$

$$\begin{aligned}\left|\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]-\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]\right| &\leq \sqrt{\frac{\mathbb{V}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}\log\left(\frac{1}{\delta}\right)} \\ &\leq \sqrt{\frac{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}\log\left(\frac{1}{\delta}\right)}.\end{aligned}$$

Here, the last inequality follows from

$$\begin{aligned}\mathbb{V}_{\boldsymbol{\gamma}}[(\ell(\tilde{\boldsymbol{\xi}})-t)^2] &= \mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\tilde{\boldsymbol{\xi}})-t)^4]-\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\tilde{\boldsymbol{\xi}})-t)^2]^2 \\ &\leq \mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\tilde{\boldsymbol{\xi}})-t)^4] \\ &\leq \mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\tilde{\boldsymbol{\xi}})-t)^2],\end{aligned}$$

where the final inequality holds because the random variable $(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2$ is supported on a subset of $[0,1]$. Next, expanding the absolute value term yields the following two cases:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]-\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2] \leq \sqrt{\frac{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}\log\left(\frac{1}{\delta}\right)} \quad \text{and} \\ &\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]-\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2] \leq \sqrt{\frac{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}\log\left(\frac{1}{\delta}\right)}.\end{aligned} \tag{16}$$

From the first case, we obtain

$$\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]-\sqrt{\frac{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}\log\left(\frac{1}{\delta}\right)} \leq \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2],$$

which is equivalent to

$$\left(\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}-\frac{1}{2}\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}\right)^2 \leq \frac{1}{4}\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}+\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2].$$

Taking square root on both sides then yields

$$\begin{aligned}\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}-\frac{1}{2}\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}} &\leq \sqrt{\frac{1}{4}\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}+\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]} \\ &\leq \frac{1}{2}\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}+\sqrt{\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]},\end{aligned} \tag{17}$$

where the last inequality follows from the relation $\sqrt{a_1+a_2} \leq \sqrt{a_1}+\sqrt{a_2}$. Next, the second case

in (16) yields

$$\begin{aligned}\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2] &\leq \mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2] + \sqrt{\frac{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}\log\left(\frac{1}{\delta}\right)} \\ &= \left(\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]} + \frac{1}{2}\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}\right)^2 - \frac{1}{4}\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)} \\ &\leq \left(\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]} + \frac{1}{2}\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}\right)^2.\end{aligned}$$

Finally, taking square root on both sides and combining with the inequality in (17), we conclude that the bound in (15) indeed holds. This completes the proof. □

Using this lemma, we obtain a bound on the error introduced by the empirical conditional standard deviation.

Proposition 1. *Fix a tolerance level $\tau > 0$. Then, we have*

$$\left|\sqrt{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]} - \sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}\right| \leq \tau + \sqrt{\frac{\log\left(\frac{1+2/\tau}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}, \tag{18}$$

with probability at least $1-\delta$.

Proof. We first show that the function $\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}$ is Lipschitz continuous in t with constant 1. Indeed, by the reverse triangle inequality, we have

$$\left|\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]} - \sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t')^2]}\right| \leq \sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t-\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})+t')^2]} = |t-t'|, \tag{19}$$

where the inequality holds because the function $\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\cdot)^2]}$ constitutes a semi-norm. One can similarly show that the function $\sqrt{\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}$ is Lipschitz continuous in t with constant 1. We next observe that

$$\sqrt{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]} = \min_{t\in[0,1]} \sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]} \quad \text{and} \quad \sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]} = \min_{t\in[0,1]} \sqrt{\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]},$$

which follows from the fact that the minimizers of these scalar optimization problems are respectively given by the mean $\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$ and the empirical mean $\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]$. Consider now a finite subset $\mathcal{T} = \{0, \tau, 2\tau, ..., 1\}$ of $[0,1]$ with cardinality $|\mathcal{T}| = 1 + 1/\tau$. Let $t^\star$ and $\hat{t}^\star$, respectively, be the minimizers of the above optimization problems over the subset $\mathcal{T}$. By the Lipschitz continuity of the objective functions, we can guarantee that

$$\left|\sqrt{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]} - \sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t^\star)^2]}\right| \leq \tau \ \text{ and } \ \left|\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]} - \sqrt{\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-\hat{t}^\star)^2]}\right| \leq \tau.$$

Thus, to ensure that the bound $\left|\sqrt{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}-\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}\right|\leq\epsilon$ holds, we simply require

$$\left|\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t^{\star})^2]}-\sqrt{\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-\hat{t}^{\star})^2]}\right|\leq\epsilon-2\tau.$$

Note that the left-hand side expression is upper bounded by the largest error

$$\max_{t\in\mathcal{T}}\left|\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}-\sqrt{\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t)^2]}\right|.$$

Thus, applying the union bound to (15) over $t\in\mathcal{T}$ yields an upper bound on left-hand side expression, as follows

$$\left|\sqrt{\mathbb{E}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-t^{\star})^2]}-\sqrt{\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})-\hat{t}^{\star})^2]}\right|\leq\sqrt{\frac{\log\left(\frac{|\mathcal{T}|}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}.$$

The result then follows by equating the right hand side with $\epsilon-2\tau$. □

We remark that the tolerance level τ can be made small without significantly increasing the square root term on the right-hand side of (18) as the latter displays merely a logarithmic dependence in τ.

Using the result in Proposition 1, we obtain a new generalization bound in view of the empirical conditional standard deviation.

Theorem 3. *Fix a tolerance level $\tau>0$. Then, for any $\boldsymbol{x}\in\mathcal{X}$, we have*

$$\left|\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]-\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]\right|\leq\left(\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}+\tau\right)\sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}+\frac{\sqrt{\log\left(\frac{2(1+2/\tau)}{\delta}\right)\log\left(\frac{2}{\delta}\right)}}{nh_n^p g(\boldsymbol{\gamma})\theta(1)},\quad(20)$$

with probability at least $1-\delta$.

Proof. The bounds in (13) and (18) yield

$$\begin{aligned}\left|\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]-\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]\right| &\leq\left(\sqrt{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}-\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}+\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}\right)\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}\\ &\leq\left(\tau+\sqrt{\frac{\log\left(\frac{1+2/\tau}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}+\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x},\tilde{\boldsymbol{\xi}})]}\right)\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}.\end{aligned}$$

The above inequality holds with probability at least $1-2\delta$, which completes the proof. □

The theorem shows that the errors introduced by replacing the conditional variance term with its empirical estimates diminish at the faster rate of $O(1/(nh_n^p))$, and become negligible when the sample size is large.

We close the section by analyzing the suboptimality bound resulting from solving the regularized problem ($\mathcal{RNW}$). To this end, we assume that the feasible set $\mathcal{X}$ is finite even though its cardinality can be exponential in the problem dimensions. Let $\hat{\boldsymbol{x}}$ be a minimizer of the regularized problem and $\boldsymbol{x}^\star$ be a minimizer of the true stochastic optimization problem ($\mathcal{S}$).

Theorem 4 (Suboptimality Bound)**.** *Fix a tolerance level $\tau > 0$. Then, for some scaling of the tuning parameter $\lambda = O\left(1/\sqrt{nh_n^p g(\boldsymbol{\gamma})}\right)$, we have*

$$\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] \leq \mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})] + \left(\sqrt{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})]} + \tau\right) \sqrt{\frac{4\log\left(\frac{6|\mathcal{X}|}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}} + \frac{2\log\left(\frac{6|\mathcal{X}|(1+2/\tau)}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}, \tag{21}$$

with probability at least $1-\delta$.

Proof. Applying the union bound to (20) over $\boldsymbol{x} \in \mathcal{X}$, we find that with probability at least $1-\delta$,

$$\left|\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})] - \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]\right| \leq \left(\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}, \tilde{\boldsymbol{\xi}})]} + \tau\right) \sqrt{\frac{\log\left(\frac{2|\mathcal{X}|}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}} + \frac{\log\left(\frac{2|\mathcal{X}|(1+2/\tau)}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)} \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

Thus, for $\boldsymbol{x} = \hat{\boldsymbol{x}}$ we get

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] &\leq \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] + \left(\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})]} + \tau\right) \sqrt{\frac{\log\left(\frac{2|\mathcal{X}|}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}} + \frac{\log\left(\frac{2|\mathcal{X}|(1+2/\tau)}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)} \\
&\leq \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})] + \left(\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})]} + \tau\right) \sqrt{\frac{\log\left(\frac{2|\mathcal{X}|}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}} + \frac{\log\left(\frac{2|\mathcal{X}|(1+2/\tau)}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)},
\end{aligned}$$

where the second inequality holds because $\boldsymbol{x}^\star$ is suboptimal for the regularized problem ($\mathcal{RNW}$). Next, applying the bound (13) for $\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})]$ and the bound (18) for $\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})]}$, we obtain

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\gamma}}[\ell(\hat{\boldsymbol{x}}, \tilde{\boldsymbol{\xi}})] \leq\ & \mathbb{E}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})] + \sqrt{\frac{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})]}{nh_n^p g(\boldsymbol{\gamma})\theta(1)} \log\left(\frac{3}{\delta}\right)} \\
&+ \left(\sqrt{\mathbb{V}_{\boldsymbol{\gamma}}[\ell(\boldsymbol{x}^\star, \tilde{\boldsymbol{\xi}})]} + 2\tau + \sqrt{\frac{\log\left(\frac{3+6/\tau}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}}\right) \sqrt{\frac{\log\left(\frac{6|\mathcal{X}|}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}} \\
&+ \frac{\log\left(\frac{6|\mathcal{X}|(1+2/\tau)}{\delta}\right)}{nh_n^p g(\boldsymbol{\gamma})\theta(1)}.
\end{aligned}$$

Finally, after performing further algebraic simplifications, we arrive at the desired bound. This completes the proof. □

Theorem 4 asserts that if there is an optimal solution $\boldsymbol{x}^\star$ of the stochastic problem ($\mathcal{S}$) which yields a cost with negligible empirical conditional variance then the regularized solution $\hat{\boldsymbol{x}}$ will converge

to this optimal solution at the faster rate of $O(1/(nh_n^p))$. Note that the bound (21) grows only logarithmically in the cardinality of the feasible set $\mathcal{X}$ and, hence, at most linearly in the dimension of the decision vector $\boldsymbol{x}$. Improved bounds with similar guarantees can be derived for other classes of cost function and feasible set by taking into account the class' Rademacher complexity and VC dimension [BM02, Vap98, SSBD14].

4 Application in portfolio management

A generic portfolio optimization problem with side information is formulated as

$$\max_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\boldsymbol{\gamma}}\left[\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x}\right], \tag{22}$$

where $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^q$ is the vector of asset returns and $\mathcal{X}$ is the set of structural constraints on the portfolio allocation vector $\boldsymbol{x} \in \mathbb{R}^q$. In this problem, the exogenous covariate vector $\tilde{\boldsymbol{\gamma}}$ may comprise the firms' market capitalizations, book-to-market ratios, past returns, and also include other market indicators such as the volatility indices and financial news indicators [BSCV09, BMD18]. In [BSCV09], the model is solved in view of linear decision rules where one seeks for the best linear policy in the exogenous covariates that maximizes the empirical return. An L_2 regularized version of the linear decision rule approximation is studied in [BMD18].

In this section, we investigate the performance of our proposed regularized NW approximation on the portfolio optimization problem (22). We first establish that the approximation is amenable to a tractable conic programming reformulation.

Proposition 2. *The regularized NW approximation*

$$\max_{\boldsymbol{x}\in\mathcal{X}} \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x}] - \lambda\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}\left[\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x}\right]} \tag{23}$$

can equivalently be reformulated as the second-order cone program

$$\begin{array}{ll} \max & \left(\displaystyle\sum_{i=1}^n p_{\boldsymbol{\gamma},i}\boldsymbol{\xi}_i^\top \boldsymbol{x}\right) - \lambda\rho \\ \text{s.t.} & \boldsymbol{x}\in\mathcal{X},\ t\in\mathbb{R} \\ & \left(\sqrt{p_{\boldsymbol{\gamma},1}}({\boldsymbol{\xi}_1}^\top \boldsymbol{x} - t), \ldots, \sqrt{p_{\boldsymbol{\gamma},n}}({\boldsymbol{\xi}_n}^\top \boldsymbol{x} - t), \rho\right) \in \mathcal{SOC}(n+1), \end{array}$$

where $p_{\boldsymbol{\gamma},i} = \frac{\mathcal{K}_h(\boldsymbol{\gamma}-\boldsymbol{\gamma}^i)}{\sum_{j=1}^n \mathcal{K}_h(\boldsymbol{\gamma}-\boldsymbol{\gamma}^j)}$, $i \in [n]$.

Proof. We first recall that $\sqrt{\hat{\mathbb{V}}_{\boldsymbol{\gamma}}[L(\tilde{\boldsymbol{\xi}})]} = \min_{t\in\mathbb{R}} \sqrt{\hat{\mathbb{E}}_{\boldsymbol{\gamma}}[(L(\tilde{\boldsymbol{\xi}}) - t)^2]}$. Thus, by introducing the aux-

iliary decision variable t, we arrive at the optimization problem

$$\begin{aligned} & \max_{\boldsymbol{x}\in\mathcal{X},t\in\mathbb{R}} \hat{\mathbb{E}}_{\boldsymbol{\gamma}}[\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x}] - \lambda\sqrt{\hat{\mathbb{E}}_{\boldsymbol{\gamma}}\left[(\tilde{\boldsymbol{\xi}}^\top \boldsymbol{x} - t)^2\right]} \\ = & \max_{\boldsymbol{x}\in\mathcal{X},t\in\mathbb{R}} \left(\sum_{i=1}^n p_{\boldsymbol{\gamma},i}{\boldsymbol{\xi}_i}^\top \boldsymbol{x}\right) - \lambda\sqrt{\sum_{i=1}^n p_{\boldsymbol{\gamma},i}({\boldsymbol{\xi}_i}^\top \boldsymbol{x} - t)^2}. \end{aligned}$$

The result then follows by introducing an epigraphical variable ρ to bring the square root term into the constraint and noting that $\|\boldsymbol{y}\| \leq \rho \Longleftrightarrow (\boldsymbol{y}, \rho) \in \mathcal{SOC}(n+1)$ for any vector $\boldsymbol{y} \in \mathbb{R}^n$ and scalar $\rho \in \mathbb{R}$. □

Example: a three-asset portfolio

We consider a portfolio optimization problem with synthetic data, which will be described in detail below. We compare our regularized NW approximation (23) with the state-of-the-art linear decision rule (LDR) formulation proposed in [BSCV09, BMD18]. The purpose of this example is to demonstrate that, even on a very simple setting, the LDR approach can fail miserably at exploiting the side information. On the other hand, the proposed regularized NW approximation is highly effective at leveraging the side information and can generate a remarkably higher average return with minimal risks.

We assume that the covariate $\tilde{\gamma} \in \mathbb{R}$ is governed by a uniform distribution on the interval $[-1, 1]$. There are three assets in total: Asset 1 and Asset 2 are risky assets, while Asset 3 is a risk-free asset. The returns of Assets 1 and 2 obey the relations

$$\tilde{\xi}_i(\tilde{\gamma}) = \frac{1}{2} - \tilde{\gamma}^2 + 0.1 \cdot \tilde{\epsilon}_i \qquad \forall i = 1, 2,$$

where the random variables $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$ are bivariate normally distributed with zero means and unit variances, and are perfectly negatively correlated. Asset 3 is assumed to have a deterministic return of 0. We adopt the standard setting in portfolio optimization where the manager does not engage in short selling and consider the problem of allocating a total wealth of \$1.

Under the above setup, one can derive the optimal expected portfolio returns with and without consideration of the side information $\tilde{\gamma}$, respectively. For the latter case, by construction, the unconditional expected returns of the risky assets are given by

$$\mathbb{E}\left[\frac{1}{2} - \tilde{\gamma}^2 + 0.1 \cdot \tilde{\epsilon}_i\right] = \frac{1}{2} - \frac{1}{3} + 0 = \frac{1}{6} \qquad \forall i = 1, 2.$$

Thus, in the absence of any side information, the optimal expected portfolio return is $1/6$, which can be obtained by allocating the entire wealth into any convex combination of the risky assets. On the other hand, suppose that the value of the covariate γ is revealed before the decision is made. In this case, the conditional expected return of each risky asset is $\frac{1}{2} - \gamma^2$. Hence, when $\gamma^2 < 1/2$, it is optimal to allocate the entire wealth into any convex combination of the risky assets; otherwise, it is optimal to allocate the entire wealth into the risk free asset. Since $\tilde{\gamma}$ follows a uniform distribution

on $[-1, 1]$, the optimal expected return of this strategy is given by

$$\int_{-\sqrt{\frac{1}{2}}}^{\sqrt{\frac{1}{2}}} \frac{1}{2}\left(\frac{1}{2} - \gamma^2\right) \, \mathrm{d}\gamma = \frac{2}{3}\left(\frac{1}{2}\right)^{3/2} \approx 0.2357.$$

Note that the expected return deteriorates by more than 40 percent if the portfolio manager ignores the side information.

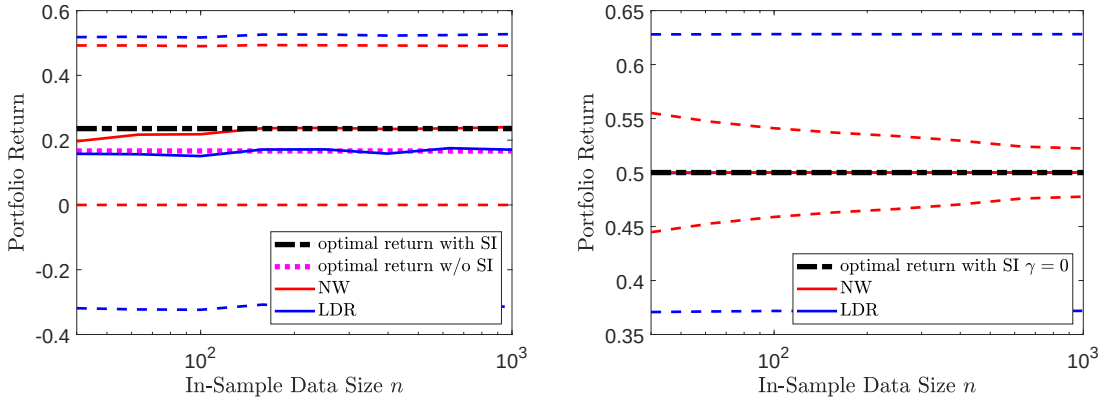
We now empirically test the proposed regularized NW approximation and the LDR formulation, and see how they perform against these optimal returns. All optimization problems are solved using Gurobi via the YALMIP interface [LÖ04] on a 4-core 3.4 GHz computer with 32 GB RAM. The dataset in the experiment is generated by taking n samples of the random vector $(\tilde{\gamma}, \tilde{\xi}_1, \tilde{\xi}_2)$. To obtain the regularized NW portfolios, we solve the SOCP formulation in Proposition 2. The LDR approach, on the other hand, seeks for the best parameters $x_1, x_2, y \in \mathbb{R}$ so that the decision of investing $\$(x_i + \gamma \cdot y)$ in Asset i, for $i = 1, 2$, and $\$(1 - x_1 - x_2 - 2\gamma \cdot y)$ in Asset 3 generates the highest empirical return. The optimal portfolio allocation thus constitutes an affine function in γ. To find these parameters, we solve the following regularized empirical maximization problem as proposed in [BMD18, BSCV09]:

$$\begin{array}{ll} \max & \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\left(\boldsymbol{\xi}_i^\top\left[\boldsymbol{x} + (\gamma_i \cdot y)\cdot \mathbf{e}\right]\right) - \lambda y^2 \\ \text{s.t.} & x_1, x_2, y \in \mathbb{R} \\ & x_1 + \gamma_i \cdot y \geq 0,\ x_2 + \gamma_i \cdot y \geq 0,\ x_1 + x_2 + 2\gamma_i \cdot y \leq 1 \qquad \forall i \in [n]. \end{array} \tag{24}$$

The constraints of this problem prohibit short selling and ensure that the total allocation does not exceed \$1.

Figure 1a shows the out-of-sample returns of the two approaches, as well as the optimal expected portfolio returns with and without consideration of the side information, respectively. We find that our proposed approach substantially outperforms LDR in terms of both return and risk. Even though the two approaches attempt to exploit the side information when generating their portfolios, the NW approach is more effective in capitalizing the information as it consistently generates higher expected returns. We also observe that the NW returns have significantly lower variability. This is not entirely surprising because the regularization term encourages a portfolio with lower standard deviation. Figure 1b depicts the out-of-sample returns for a fixed covariate $\gamma = 0$. In this case, the conditional expected return of each risky asset is 0.5 and investing in any convex combination of the two risky assets yields the optimal expected portfolio return. Since Asset 1 and Asset 2 have perfect negative correlation, the NW approach tends to allocate an equal weight to both assets so that the individual noise terms $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$ are neutralized in the resulting portfolio.

As expected, the returns of the NW approximation converge fast to the best expected portfolio return as the data size grows. On the other hand, we observe that LDR disappointingly performs as if it were oblivious to the side information, even with large data size. This phenomenon can be



(a) γ's are sampled from uniform distribution within $[-1, 1]$

(b) $\gamma = 0.0$

Figure 1: Out-of-sample portfolio returns of different approaches over 1000 γ's for each n. The black dot line is the optimal expected return with the side information γ given. The pink dot line is the optimal expected return without considering side information. The red and blue soild lines are the average returns of our proposed model and the LDR formulation, respectively. The two dot lines for each color record the 10th and 90th percentile returns of the corresponding approach.

explained analytically as follows. For any fixed parameters x_1, x_2, and y, the expected portfolio return is given by

$$\begin{aligned} & \mathbb{E}[\tilde{\xi}_1(\tilde{\gamma})(x_1 + \tilde{\gamma} \cdot y) + \tilde{\xi}_2(\tilde{\gamma})(x_2 + \tilde{\gamma} \cdot y)] \\ = \; & \mathbb{E}\left[\left(\tfrac{1}{2} - \tilde{\gamma}^2 + 0.1 \cdot \tilde{\epsilon}_1\right)(x_1 + \tilde{\gamma} \cdot y) + \left(\tfrac{1}{2} - \tilde{\gamma}^2 + 0.1 \cdot \tilde{\epsilon}_2\right)(x_2 + \tilde{\gamma} \cdot y)\right] \\ = \; & \mathbb{E}\left[\left(\tfrac{1}{2} - \tilde{\gamma}^2\right)(x_1 + x_2) + 2\left(\tfrac{1}{2}\tilde{\gamma} \cdot y - \tilde{\gamma}^3 \cdot y\right)\right] \\ = \; & \left(\tfrac{1}{2} - \tfrac{1}{3}\right)(x_1 + x_2) + 0 \\ = \; & \frac{x_1 + x_2}{6}, \end{aligned}$$

where the third equality holds because the random variables $\tilde{\epsilon}_1$ and $\tilde{\epsilon}_2$ are independent of $\tilde{\gamma}$ and have mean zero, while the penultimate equality follows from the identities $\mathbb{E}[\tilde{\gamma}^2] = 1/3$ and $\mathbb{E}[\tilde{\gamma}] = \mathbb{E}[\tilde{\gamma}^3] = 0$. Since the constraint $x_1 + x_2 \leq 1$ is imposed in the formulation, the LDR approach will never generate an expected portfolio return greater than $1/6$. This result affirms our observation that LDR indeed performs as poorly as the model that disregards the side information.

5 Concluding remarks

The NW approximation has recently garnered an increasing interest due to its significance in the context of decision-making under uncertainty with side information. The scheme, however, has so far resisted any sensible result about its out-of-sample performance. In this paper, we established for the first time a complete, comprehensive theoretical result on the performance guarantees of the approximation. The new result inspired us to design a novel regularization scheme that can

better mitigate the overfitting effects. In contrast to the popular L_2 regularization scheme which attempts to minimize the norm of the decision vector and may pointlessly encourage an optimal solution that is close to the origin, our proposed regularization scheme is directly constructed using the conditional standard deviation term appearing in the theoretical bounds and can faithfully prioritize an optimal solution that generalizes well. In the future, it would be interesting to establish a connection between the new regularization scheme and the distributionally robust optimization paradigm. It is also imperative to extend the model to the multi-stage setting, and devise a tractable solution procedure with similar performance guarantees for the dynamic stochastic optimization problems.

Acknowledgements. Grani A. Hanasusanto is supported by the National Science Foundation grant no. 1752125.

References

[BK14] D. Bertsimas and N. Kallus. From predictive to prescriptive analytics. *arXiv preprint arXiv:1402.5481*, 2014.

[BM02] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[BMD18] T. Bazier-Mattea and E. Delage. Generalization bounds for regularized portfolio selection with market side information. *Optimization Online*, 2018.

[BR18] G.-Y. Ban and C. Rudin. The big data newsvendor: Practical insights from machine learning. *Operations Research*, 2018.

[BSCV09] M. W. Brandt, P. Santa-Clara, and R. Valkanov. Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns. *The Review of Financial Studies*, 22(9):3411–3447, 2009.

[BVP17] D. Bertsimas and B. Van Parys. Bootstrap robust prescriptive analytics. *arXiv preprint arXiv:1711.09974*, 2017.

[DZ98] A. Dembo and O. Zeitouni. Large deviations techniques and applications. *Applications of Mathematics*, 38, 1998.

[EL03] P. Eichelsbacher and M. Löwe. Moderate deviations for iid random variables. *ESAIM: Probability and Statistics*, 7:209–218, 2003.

[GKKW06] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.

[HK13] G. A. Hanasusanto and D. Kuhn. Robust data-driven dynamic programming. In *Advances in Neural Information Processing Systems*, pages 827–835, 2013.

[HPB10] L. Hannah, W. Powell, and D. Blei. Nonparametric density estimation for stochastic optimization with an observable state variable. In *Advances in Neural Information Processing Systems*, pages 820–828, 2010.

[KSH02] A. J. Kleywegt, A. Shapiro, and T. Homem-de-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

[LÖ4] J. Löfberg. YALMIP : A toolbox for modeling and optimization in MATLAB. In *Proceedings of the CACSD Conference*, 2004.

[MPT08] A. Mokkadem, M. Pelletier, and B. Thiam. Large and moderate deviations principles for kernel estimators of the multivariate regression. *Mathematical Methods of Statistics*, 17(2):146–172, 2008.

[Nad64] E. A. Nadaraya. On estimating regression. *Theory of Probability & its Applications*, 9(1):141–142, 1964.

[SDR09] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[Sil86] B. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26. Chapman & Hall/CRC, 1986.

[SSBD14] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[Vap98] V. N. Vapnik. Statistical learning theory. *A Wiley-Interscience Publication*, 1998.

[Wat64] G. S. Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.