2019

# Computational Aspects of Bayesian Solution Estimators in Stochastic Optimization

Danial Davarnia
*Iowa State University*, davarnia@iastate.edu

Burak Kocuk
*Sabanci University*

Gerard Cornuejols
*Carnegie Mellon University*

# Computational Aspects of Bayesian Solution Estimators in Stochastic Optimization


## Abstract

We study a class of stochastic programs where some of the elements in the objective function are random, and their probability distribution has unknown parameters. The goal is to find a good estimate for the optimal solution of the stochastic program using data sampled from the distribution of the random elements. We investigate two common optimization criteria for evaluating the quality of a solution estimator, one based on the difference in objective values, and the other based on the Euclidean distance between solutions. We use risk as the expected value of such criteria over the sample space. Under a Bayesian framework, where a prior distribution is assumed for the unknown parameters, two natural estimation-optimization strategies arise. A separate scheme first finds an estimator for the unknown parameters, and then uses this estimator in the optimization problem. A joint scheme combines the estimation and optimization steps by directly adjusting the distribution in the stochastic program. We analyze the risk difference between the solutions obtained from these two schemes for several classes of stochastic programs, while providing insight on the computational effort to solve these problems.



## Keywords

Stochastic optimization, Bayesian inference, Statistical estimation, Solution estimators


## Disciplines

Operations Research, Systems Engineering and Industrial Engineering | Statistics and Probability

## Comments

This is a pre-print of the article Davarnia, Danial, Burak Kocuk, and Gérard Cornuéjols. "Computational Aspects of Bayesian Solution Estimators in Stochastic Optimization." (2019). Posted with permission.

**Submitted to *INFORMS Journal on Optimization***
**manuscript**



# Computational Aspects of Bayesian Solution Estimators in Stochastic Optimization

Danial Davarnia
Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA, USA, davarnia@iastate.edu

Burak Kocuk
Industrial Engineering Program, Sabancı University, Istanbul, Turkey, burakkocuk@sabanciuniv.edu

Gérard Cornuéjols
Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA, gc0v@andrew.cmu.edu

We study a class of stochastic programs where some of the elements in the objective function are random, and their probability distribution has unknown parameters. The goal is to find a good estimate for the optimal solution of the stochastic program using data sampled from the distribution of the random elements. We investigate two common optimization criteria for evaluating the quality of a solution estimator, one based on the difference in objective values, and the other based on the Euclidean distance between solutions. We use *risk* as the expected value of such criteria over the sample space. Under a Bayesian framework, where a prior distribution is assumed for the unknown parameters, two natural estimation-optimization strategies arise. A *separate* scheme first finds an estimator for the unknown parameters, and then uses this estimator in the optimization problem. A *joint* scheme combines the estimation and optimization steps by directly adjusting the distribution in the stochastic program. We analyze the risk difference between the solutions obtained from these two schemes for several classes of stochastic programs, while providing insight on the computational effort to solve these problems.

*Key words*: Stochastic optimization, Bayesian inference, Statistical estimation, Solution estimators.

## 1. Introduction

The interaction between data and optimization can be complex and sometimes counterintuitive. As an example, consider portfolio optimization: Construct a portfolio of assets that maximizes expected return over some future period, subject to a constraint on risk (Markowitz 1959). A major issue in practice is the estimation of the expected returns of the individual assets. The theory of efficient markets tells us that the latest data contain the best available information. Surprisingly, Jorion (1986) recommends not to use the data directly to estimate the expected asset returns,

but instead to shrink the data on individual assets towards a "grand average" involving all the other assets as well. This seems counterintuitive but it works well in practice. This "paradox" can be traced back to the work of Stein (1956). Such results are best understood under a Bayesian framework. In this paper we focus on the computational implications of parameter estimation in the context of stochastic optimization.

Consider the following stochastic program

$$\max_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi},\boldsymbol{x})]. \tag{1}$$

In (1), the vector $\boldsymbol{\xi}\in\mathbb{R}^n$ represents random variables with joint probability density function $g(\boldsymbol{\xi}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes a vector of parameters. Vector $\boldsymbol{x}\in\mathbb{R}^m$ represents decision variables that belong to a closed set $\mathcal{X}\subseteq\mathbb{R}^m$. Function $f(\boldsymbol{\xi},\boldsymbol{x}):\mathbb{R}^{n+m}\to\mathbb{R}$ contains both random variables and decision variables, and its expectation with respect to the distribution $g(\boldsymbol{\xi}|\boldsymbol{\theta})$ forms the objective function of the stochastic program. We assume that this expectation is finite for all $\boldsymbol{x}\in\mathcal{X}$. In our introductory portfolio example, the asset returns $\boldsymbol{\xi}$ could have a multivariate normal distribution $g$ with mean $\boldsymbol{\theta}$ and known covariance matrix, the portfolio weights form the decision vector $\boldsymbol{x}$, and the portfolio return is $f(\boldsymbol{\xi},\boldsymbol{x})=\boldsymbol{\xi}^\top\boldsymbol{x}$.

If the parameters $\boldsymbol{\theta}$ are known, problem (1) reduces to a classical stochastic program with an optimal solution $\boldsymbol{x}^*(\boldsymbol{\theta})$ and optimal value $\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi},\boldsymbol{x}^*(\boldsymbol{\theta}))]$. In *parametric* stochastic programs, it is assumed that $\boldsymbol{\theta}$ is not fully known. In such situations, $T\geq 1$ independent observations of the random variables $\boldsymbol{\xi}$ drawn from the distribution $g(\boldsymbol{\xi}|\boldsymbol{\theta})$ are used to estimate the unknown parameters. The question is how to use these data to find a *good* estimator for the true optimal solution $\boldsymbol{x}^*(\boldsymbol{\theta})$ of (1)?

Parametric stochastic programs encompass numerous applications; see Birge and Louveaux (2011) for applications in optimization and Donti et al. (2017) for an application in machine learning. Various solution techniques have been proposed in the literature. Examples are Monte Carlo, bootstrapping and scenario reduction techniques such as sample-average approximation (SAA); see Shapiro (2003). Jiang and Shanbhag (2016) consider a setup where the parameter may be obtained through the solution of a learning problem. See also Wang et al. (2017).

One of the main tools in studying parametric stochastic problems is point estimation, which originates in statistical inference and decision theory. We refer the interested reader to Ferguson (1967) or Lehmann and Casella (1998) and many references therein. Estimation plays a central role in constructing optimization models that involve uncertainty. For instance, the theory of minimaxity in estimation is closely related to robust optimization. We refer the reader to Ben-Tal et al. (2009), Bertsimas et al. (2011) for an introduction to the theory and applications of

robust optimization, and to Lim et al. (2006) and references therein for a detailed account of the connection between robust optimization and minimaxity.

In this paper, we focus on stochastic optimization and its connection to statistical estimation theory under a Bayesian setting, where a prior distribution is assumed for the unknown parameters; see Kadane (2011) for an overview of basics in the Bayesian framework. This is a common framework when modeling parametric stochastic programs; see Lim et al. (2006). For the portfolio optimization problem introduced earlier, this setting implies that the returns have a certain distribution whose mean is random and follows a prior distribution. There are two natural schemes to find a solution estimator in this setting. In the first scheme, the parameter estimation is performed separately, and its estimator is used as an input for the optimization problem. In the second scheme, the estimation process is incorporated within the optimization step to obtain a solution.

While several bodies of work provide comparisons between the "true" stochastic optimization problem and its approximate "deterministic" counterpart for classical stochastic programs (see Kleywegt and Shapiro 2004), sometimes framed under the "separate" versus "joint" estimation-optimization framework (see e.g., Donti et al. 2017, Levine et al. 2016, Thomas et al. 2006), a detailed analysis of this nature is lacking in parametric stochastic optimization under the Bayesian framework. We provide such an analysis for the above schemes from two angles: the quality of the solution estimators obtained by each scheme, and the computational efficiency of performing them. In particular, we show how different problem structures and probability distributions affect the quality and efficiency of the separate and joint schemes, by providing theoretical and computational results.

Our study is organized as follows. In Section 2, we formally define the key problem settings: risk criteria, and the separate and joint estimation-optimization schemes. In Section 3, we illustrate the tradeoff between accuracy and computational efficiency by presenting experiments on two applications: a portfolio optimization problem and a stochastic geometric program. We observe that the separate and joint schemes can lead to problems with different computational complexity and solution stability. In Section 4, we identify conditions under which the two schemes yield the same solutions, and provide examples with an arbitrarily large risk difference when these conditions do not hold. In Section 5, we study a class of stochastic piecewise linear programs under our two schemes. This class of stochastic programs generalizes the newsvendor and median problems, and models applications in education, psychology and artificial intelligence. We derive explicit bounds on the risk difference between the solution estimators of the separate and joint schemes for various probability distributions. Through a higher dimensional extension, this class also describes the "rectifier" used as an activation function for deep neural networks in machine learning. We conclude the section by performing computational experiments on a core structure of such models. Section 6 contains concluding remarks.

## 2. Loss function and risk

Consider the stochastic program (1). The form of the distribution $g(\boldsymbol{\xi}|\boldsymbol{\theta})$ is known but the parameters $\boldsymbol{\theta}$ are unknown. However, we have $T \geq 1$ observations sampled from this distribution. To simplify notation, we will assume a single vector $\bar{\boldsymbol{\xi}}$ of observations, i.e., $T = 1$. This is done without loss of generality since in our Bayesian context the samples are used to derive posterior distributions, which are readily obtainable for any size $T$.

Since the true optimal solution $\boldsymbol{x}^*(\boldsymbol{\theta})$ is unknown (as $\boldsymbol{\theta}$ is unknown), we estimate it with a *solution estimator* $\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}) \in \mathcal{X}$ as a function of the observation $\bar{\boldsymbol{\xi}}$. There are many choices for a solution estimator, hence we need a suitable criterion to evaluate its performance. We use *risk*, a popular criterion in statistical inference. To this end, we define two natural *loss* functions, one based on the difference in objective values and the other based on the distance between solutions.

### 2.1. Loss as the difference in objective values

The *loss function* $\mathcal{L}^L$ is defined as the difference between the objective function values of the solution estimator $\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})$ and the true optimal solution $\boldsymbol{x}^*(\boldsymbol{\theta})$,

$$\mathcal{L}^L\left(\boldsymbol{x}^*(\boldsymbol{\theta}), \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})\right) = \mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi}, \boldsymbol{x}^*(\boldsymbol{\theta}))] - \mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi}, \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))], \tag{2}$$

where the expectations are taken with respect to the distribution $g(\boldsymbol{\xi}|\boldsymbol{\theta})$ of $\boldsymbol{\xi}$ given the true $\boldsymbol{\theta}$. The superscript represents that the loss is *linear* in objective function values. Note that $\mathcal{L}^L\left(\boldsymbol{x}^*(\boldsymbol{\theta}), \boldsymbol{x}^*(\boldsymbol{\theta})\right) = 0$ and that $\mathcal{L}^L\left(\boldsymbol{x}^*(\boldsymbol{\theta}), \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})\right) \geq 0$ for any $\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}) \in \mathcal{X}$ as it is feasible to (1), which has $\boldsymbol{x}^*(\boldsymbol{\theta})$ as an optimal solution.

We evaluate the loss under a popular framework in the Bayesian school, where the unknown parameters $\boldsymbol{\theta}$ are assumed to be random with a prior distribution $\pi(\boldsymbol{\theta})$. Since $\mathcal{L}^L\left(\boldsymbol{x}^*(\boldsymbol{\theta}), \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})\right)$ is a function of both the unknown parameters $\boldsymbol{\theta}$ and the observation $\bar{\boldsymbol{\xi}}$, it is a random quantity. To obtain a measure of overall performance, a *risk* is defined as

$$\mathcal{R}^L\left(\boldsymbol{x}^*(\boldsymbol{\theta}), \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})\right) = \mathbb{E}_\theta \mathbb{E}_{\bar{\xi}|\theta}\left[\mathcal{L}^L\left(\boldsymbol{x}^*(\boldsymbol{\theta}), \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})\right)\right], \tag{3}$$

where the outer expectation is taken with respect to the prior distribution $\pi(\boldsymbol{\theta})$ of $\boldsymbol{\theta}$, and the inner expectation is taken with respect to the likelihood distribution $g(\bar{\boldsymbol{\xi}}|\boldsymbol{\theta})$.

A *best* solution estimator is defined as one that achieves minimum risk. It is called a *Bayes solution estimator*. The proof of the following proposition as well as some others in this section follow from standard techniques in Bayesian analysis. We include these proofs in Appendix A for the sake of completeness.

PROPOSITION 1. *Consider stochastic program* (1)*. Assume that* $\boldsymbol{\xi} \sim g(\boldsymbol{\xi}|\boldsymbol{\theta})$ *and* $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$*. Then, any solution estimator* $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}})$ *that solves*

$$\max_{\boldsymbol{x}\in\mathcal{X}} \ \mathbb{E}_{\theta|\bar{\xi}}\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi},\boldsymbol{x})] \tag{4}$$

*is a Bayes solution estimator under the loss function* (2)*, where the inner expectation is taken with respect to the likelihood distribution* $g(\boldsymbol{\xi}|\boldsymbol{\theta})$*, and the outer expectation is taken with respect to the posterior distribution* $\Pi(\boldsymbol{\theta}|\bar{\boldsymbol{\xi}})$ *of the parameters* $\boldsymbol{\theta}$ *given the observation* $\bar{\boldsymbol{\xi}}$*.*

The Bayes solution estimator $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}})$ found in Proposition 1 minimizes risk for the loss function (2) in a joint estimation and optimization (Joint-EO) scheme. The superscripts in $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}})$ represent the *Joint* scheme with respect to the *Linear* loss (2).

We next give another representation of the Joint-EO problem (4) based on the expectation with respect to a specific marginal distribution called posterior predictive.

DEFINITION 1. The *posterior predictive* distribution of $\boldsymbol{\xi}$ given $\bar{\boldsymbol{\xi}}$ is defined as $h(\boldsymbol{\xi}|\bar{\boldsymbol{\xi}}) = \int_\theta g(\boldsymbol{\xi}|\boldsymbol{\theta})\Pi(\boldsymbol{\theta}|\bar{\boldsymbol{\xi}})d\boldsymbol{\theta}$, which is obtained by marginalizing the distribution of $\boldsymbol{\xi}$ given $\boldsymbol{\theta}$ over the posterior distribution of $\boldsymbol{\theta}$ given observation $\bar{\boldsymbol{\xi}}$.

PROPOSITION 2. *Assume that* $\boldsymbol{\xi} \sim g(\boldsymbol{\xi}|\boldsymbol{\theta})$*,* $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$*, and that* $\mathbb{E}_{\xi|\bar{\xi}}[|f(\boldsymbol{\xi},\boldsymbol{x})|]$ *is finite for any* $\boldsymbol{x} \in \mathcal{X}$*. Then, any solution estimator* $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}})$ *that solves*

$$\max_{\boldsymbol{x}\in\mathcal{X}} \ \mathbb{E}_{\xi|\bar{\xi}}[f(\boldsymbol{\xi},\boldsymbol{x})], \tag{5}$$

*is a Bayes solution estimator under the loss function* (2)*, where the expectation is taken with respect to the posterior predictive distribution* $h(\boldsymbol{\xi}|\bar{\boldsymbol{\xi}})$*.*

Computing $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}})$ can be extremely challenging in practice. An alternative, which can be sub-optimal in risk, is to apply a separate estimation and optimization (Separate-EO) scheme. First, a Bayes estimator $\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{\xi}})$ for $\boldsymbol{\theta}$ is computed under the squared error loss, then this estimator is used in place of $\boldsymbol{\theta}$ in (1) to obtain an optimal solution $\hat{\boldsymbol{x}}^S(\bar{\boldsymbol{\xi}})$ that serves as a solution estimator for the true optimal solution $\boldsymbol{x}^*(\boldsymbol{\theta})$. The superscript in $\hat{\boldsymbol{x}}^S(\bar{\boldsymbol{\xi}})$ represents the *Separate* scheme.

It can be shown that the Bayes estimator of $\boldsymbol{\theta}$ under the squared error loss $\mathcal{L}^Q(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}(\bar{\boldsymbol{\xi}})) = ||\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}(\bar{\boldsymbol{\xi}})||^2$ is $\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{\xi}}) = \mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{\theta}]$, where the expectation is taken with respect to the posterior distribution $\Pi(\boldsymbol{\theta}|\bar{\boldsymbol{\xi}})$. When the likelihood and prior belong to the *conjugate* family, the posterior mean is representable as a convex combination between the prior mean and the observation (maximum likelihood estimator for multiple observations); see Diaconis and Ylvisaker (1979). Such estimators are sometimes referred to as *shrinkage* estimators as they shrink the maximum likelihood estimator towards another vector. Jorion (1986)'s shrinkage mentioned in the introduction can be viewed in

this light. We refer the interested reader to Davarnia and Cornuéjols (2017) and Basu et al. (2019) for a non-Bayesian perspective.

Proposition 2 allows for a direct comparison between the Separate-EO and the Joint-EO problems. For the Separate-EO, we solve (1) where the expectation is taken with respect to the distribution $g(\boldsymbol{\xi}|\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{\xi}}))$. For the Joint-EO under the linear loss (2), we solve (5) where the expectation is taken with respect to the posterior predictive distribution $h(\boldsymbol{\xi}|\bar{\boldsymbol{\xi}})$. Therefore, the difference between these two estimation schemes stems from the difference between the distributions with respect to which the expectation of $f(\boldsymbol{\xi}, \boldsymbol{x})$ is computed.

### 2.2. Loss as the distance between solutions

Since an optimization problem can have multiple optimal solutions, we let $\mathcal{D}(W, v) = \min_{w \in W} ||w - v||$, where $W \subseteq \mathbb{R}^n$ and $v \in \mathbb{R}^n$. We define the loss function

$$\mathcal{L}^Q\left(\mathcal{X}^*(\boldsymbol{\theta}), \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})\right) = \mathcal{D}^2\left(\boldsymbol{x}^*(\boldsymbol{\theta}), \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})\right), \tag{6}$$

where $\mathcal{X}^*(\boldsymbol{\theta})$ denotes the set of optimal solutions of (1) when $\boldsymbol{\theta}$ is known. The superscript indicates that the loss is *quadratic* in solutions. We define risk similarly to (3) as

$$\mathcal{R}^Q\left(\boldsymbol{x}^*(\boldsymbol{\theta}), \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})\right) = \mathbb{E}_\theta \mathbb{E}_{\bar{\xi}|\theta}\left[\mathcal{L}^Q\left(\boldsymbol{x}^*(\boldsymbol{\theta}), \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})\right)\right]. \tag{7}$$

Note that the Separate-EO method does not incorporate information about the optimization problem in the estimation step. Hence, the change of the loss function does not affect this method, and we obtain the same solution estimator $\hat{\boldsymbol{x}}^S(\bar{\boldsymbol{\xi}})$ as given in Section 2.1. For the Joint-EO method, however, the underlying optimization problem is different and therefore a different definition for solution estimators is necessary.

Proposition 3. *Assume that $\boldsymbol{\xi} \sim g(\boldsymbol{\xi}|\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$. Then, any solution estimator $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}})$ that solves*

$$\min_{\boldsymbol{x} \in \mathcal{X}} \ \mathbb{E}_{\theta|\bar{\xi}}\left[\mathcal{D}^2(\boldsymbol{x}^*(\boldsymbol{\theta}), \boldsymbol{x})\right], \tag{8}$$

*is a Bayes solution estimator under the loss function* (6)*, where the expectation is taken with respect to the posterior distribution $\Pi(\boldsymbol{\theta}|\bar{\boldsymbol{\xi}})$.*

We note that the problem (8) is indeed a minimization problem as opposed to the problems (4) and (5).

The superscript in $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}})$ indicates the *Joint* method under the *Quadratic* loss function (6). In the sequel, the expectation operator $\mathbb{E}[\cdot]$ applied on a vector implies component-wise expectation. The following result shows that $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}})$ can be obtained as the projection of $\mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\theta})]$ onto $\mathcal{X}$, when (1) has a unique optimal solution.

COROLLARY 1. *Assume that* (1) *has a unique optimal solution* $\boldsymbol{x}^*(\boldsymbol{\theta})$ *for any given* $\boldsymbol{\theta}$*. Then, the Bayes solution estimator of* (8) *under the loss function* (6) *is obtained as the optimal solution of*

$$\min_{\boldsymbol{x}\in\mathcal{X}} \ ||\mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\theta})] - \boldsymbol{x}||^2. \tag{9}$$

*Further, when* $\mathcal{X}$ *is convex, the Bayes solution estimator is* $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}}) = \mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\theta})]$.

We note that computing $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}})$ can be more challenging than $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}})$, as the former requires an explicit derivation of the optimal solution set $\boldsymbol{x}^*(\boldsymbol{\theta})$ which can be very hard for general problem structures.

## 3. Trade-off between accuracy and computational efficiency

As established in Section 2, the Joint-EO method yields the best solution estimator. On the other hand, the Separate-EO method may sometimes be computationally more efficient. In this section, we present two applications that illustrate contrasting points on this trade-off spectrum.

The first application is a portfolio construction model, and we show that the Joint-EO method dominates the Separate-EO method in both computation time and solution quality. For the second application, we investigate a stochastic geometric program. Here the Joint-EO method is computationally expensive and does not yield stable solutions, whereas the Separate-EO method is tractable and exhibits a robust performance.

### 3.1. Application I: Portfolio construction

Consider an investor with unit initial wealth. She can invest in a combination of $n$ stocks whose random return is denoted by the vector $\boldsymbol{\xi}$. We assume that random returns have a joint distribution $g(\boldsymbol{\xi}|\boldsymbol{\mu})$ with unknown mean $\boldsymbol{\mu}$. We further assume that a prior for the unknown parameters, denoted by $\pi(\boldsymbol{\mu})$, is available and a set of observations $\bar{\boldsymbol{\xi}}^1, \ldots, \bar{\boldsymbol{\xi}}^T$ are drawn from the likelihood distribution.

We assume that the utility function of the investor is given as $u : [0, \infty) \to \mathbb{R}$ and the aim is to maximize the expected utility. Denoting the portfolio weights by $\boldsymbol{x}$, we formulate this problem as

$$\max_{\boldsymbol{x}\in\Delta^n} \mathbb{E}_{\xi|\mu}[u(1 + \boldsymbol{\xi}^\top \boldsymbol{x})], \tag{10}$$

where $\Delta^n := \{\boldsymbol{x} \in \mathbb{R}^n_+ : \ \sum_{j=1}^n x_j = 1\}$. Recall that the Separate-EO estimator can be obtained as an optimal solution of

$$\max_{\boldsymbol{x}\in\Delta^n} \mathbb{E}_{\xi|\mu^B(\bar{\xi})}[u(1 + \boldsymbol{\xi}^\top \boldsymbol{x})], \tag{11}$$

where $\boldsymbol{\mu}^B(\bar{\boldsymbol{\xi}})$ is the Bayes estimator of $\boldsymbol{\mu}$, whereas the Joint-EO estimator under the linear loss can be computed as an optimal solution of

$$\max_{\boldsymbol{x}\in\Delta^n} \mathbb{E}_{\xi|\bar{\xi}}[u(1 + \boldsymbol{\xi}^\top \boldsymbol{x})]. \tag{12}$$

Now, consider a situation where the likelihood and prior distributions are multivariate normal with known covariance matrices and prior mean, that is, $\boldsymbol{\xi}|\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})$. Then, we obtain the posterior and predictive distributions respectively as $\boldsymbol{\mu}|\bar{\boldsymbol{\xi}} \sim \mathcal{N}(\boldsymbol{\mu_0'}, \boldsymbol{\Sigma_0'})$ and $\boldsymbol{\xi}|\bar{\boldsymbol{\xi}} \sim \mathcal{N}(\boldsymbol{\mu_0'}, \boldsymbol{\Sigma_0'} + \boldsymbol{\Sigma})$, where $\bar{\boldsymbol{\xi}}$ is the sample average, $\boldsymbol{\Sigma_0'} := (\boldsymbol{\Sigma_0}^{-1} + m\boldsymbol{\Sigma}^{-1})^{-1}$ and $\boldsymbol{\mu_0'} := \boldsymbol{\Sigma_0'}(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0} + m\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{\xi}})$. Here, $m$ is the sample size and $\boldsymbol{\mu_0'}$ is the posterior mean (hence, the Bayes estimator for $\mu$).

Let the utility function be of the popular exponential form, $u(\omega) := 1 - e^{-\lambda\omega}$ for some given $\lambda > 0$. It follows that $1 + \boldsymbol{\xi}^\top \boldsymbol{x}|\boldsymbol{\mu_0'} \sim \mathcal{N}(1 + \boldsymbol{x}^\top \boldsymbol{\mu_0'}, \boldsymbol{x}^\top \boldsymbol{\Sigma_0'} \boldsymbol{x})$ and $1 + \boldsymbol{\xi}^\top \boldsymbol{x}|\bar{\boldsymbol{\xi}} \sim \mathcal{N}(1 + \boldsymbol{x}^\top \boldsymbol{\mu_0'}, \boldsymbol{x}^\top (\boldsymbol{\Sigma_0'} + \boldsymbol{\Sigma})\boldsymbol{x})$. Using the moment-generating function of the normal distribution, we can obtain the Separate-EO estimator as an optimal solution of

$$\max_{\boldsymbol{x} \in \Delta^n} \left\{ 1 - e^{-\lambda\left(1 + \boldsymbol{x}^\top \boldsymbol{\mu_0'}\right) + \frac{1}{2}\lambda^2 \left(\boldsymbol{x}^\top \boldsymbol{\Sigma_0'} \boldsymbol{x}\right)} \right\},$$

while the Joint-EO estimator under the linear loss can be computed as an optimal solution of

$$\max_{\boldsymbol{x} \in \Delta^n} \left\{ 1 - e^{-\lambda\left(1 + \boldsymbol{x}^\top \boldsymbol{\mu_0'}\right) + \frac{1}{2}\lambda^2 \left(\boldsymbol{x}^\top (\boldsymbol{\Sigma_0'} + \boldsymbol{\Sigma}) \boldsymbol{x}\right)} \right\}.$$

In fact, since the exponential function is monotone, the above problems can be converted to the quadratic programs

$$\max_{\boldsymbol{x} \in \Delta^n} \left\{ \left(\boldsymbol{x}^\top \boldsymbol{\mu_0'}\right) - \frac{1}{2}\lambda \left(\boldsymbol{x}^\top \boldsymbol{\Sigma_0'} \boldsymbol{x}\right) \right\}, \tag{13}$$

and

$$\max_{\boldsymbol{x} \in \Delta^n} \left\{ \left(\boldsymbol{x}^\top \boldsymbol{\mu_0'}\right) - \frac{1}{2}\lambda \left(\boldsymbol{x}^\top (\boldsymbol{\Sigma_0'} + \boldsymbol{\Sigma}) \boldsymbol{x}\right) \right\}, \tag{14}$$

respectively.

To compare the empirical performance of the Separate-EO and Joint-EO solution estimators, we use a real dataset from Kocuk and Cornuéjols (2018) based on 11 sectors in the Standard & Poor's 500 index spanning 360 months. In particular, let $\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^{360}$ be the sector returns, and $\boldsymbol{w}^1, \ldots, \boldsymbol{w}^{360}$ be the weights of the market portfolio constructed based on the market capitalization of each sector. We use the Black-Litterman approach to construct a prior distribution; see Black and Litterman (1991) and Black and Litterman (1992) for the original derivation, and Bertsimas et al. (2012) for a modern interpretation.

Let $m$ be a fixed integer (such as 180) denoting the number of data points used in the estimation. For each $v = 181, \ldots, 360$, we conduct the following experiment:

- We compute the sample average and sample covariance $\hat{\boldsymbol{\mu}}^{v-1}$ and $\hat{\boldsymbol{\Sigma}}^{v-1}$ of the return values $\boldsymbol{\xi}^{v-m}, \ldots, \boldsymbol{\xi}^{v-1}$.
- Using the Black-Litterman approach, we set the prior mean as $\lambda \hat{\boldsymbol{\Sigma}}^{v-1} \boldsymbol{w}^{v-1}$ and prior covariance as $\tau \hat{\boldsymbol{\Sigma}}^{v-1}$, for some $\tau > 0$. Here, small (large) $\tau$ indicates strong (weak) prior while small (large) $\lambda$ represents low (high) risk aversion.

- We set the observed data mean as $\hat{\boldsymbol{\mu}}^{v-1}$ and likelihood covariance as $\hat{\boldsymbol{\Sigma}}^{v-1}$.
- We solve problems (13) and (14) to obtain the Separate-EO estimator $\hat{\boldsymbol{x}}^S$ and the Joint-EO estimator $\hat{\boldsymbol{x}}^{J,L}$.
- We compute the realized portfolio returns as $r_v^S := \boldsymbol{\xi}^{v\top}\hat{\boldsymbol{x}}^S$ and $r_v^{J,L} := \boldsymbol{\xi}^{v\top}\hat{\boldsymbol{x}}^{J,L}$.
- Finally, we compute some performance measures such as average, standard deviation, 1% Conditional Value-at-Risk (CVaR), that is, the average of the worst 1% of the realizations, and the ratio of the average to standard deviation and 1% CVaR given the realized returns $r_{181}^S, \ldots, r_{360}^S$ and $r_{181}^{J,L}, \ldots, r_{360}^{J,L}$.

The results with varying values of risk aversion $\lambda$ and prior confidence $\tau$ are given in Figure 1. We observe that the performance of the Joint-EO estimator is superior to that of the Separate-EO estimator in terms of all the measures considered. We also observe that the performance of the Separate-EO is fluctuating based on the specific choices of $\lambda$ and $\tau$ parameters whereas the performance of the Joint-EO is consistent. Due to our choice of the conjugate families, the computational effort of the separate and joint schemes are similar as the likelihood and posterior predictive are both Normal. In addition, the accuracy of the joint method is uniformly higher than the separate method. Next, we present another application that exhibits an opposite performance.

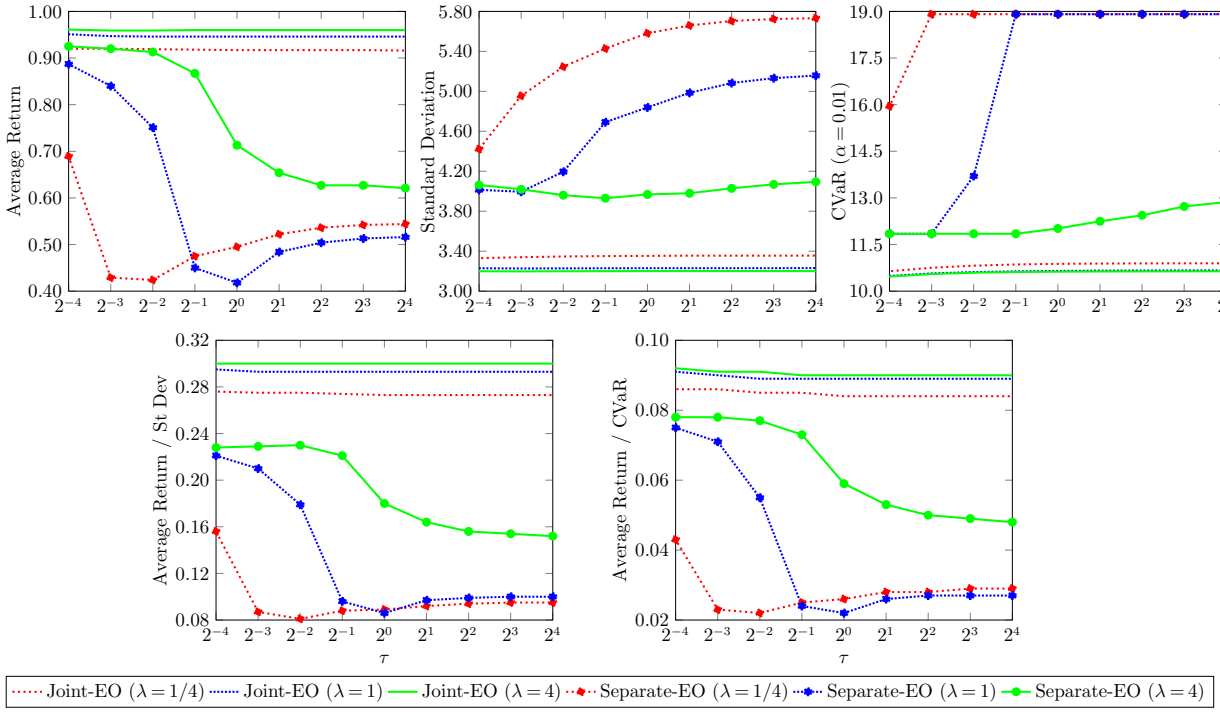


**Figure 1** **Empirical performance comparison of Separate-EO and Joint-EO solution estimators ($m = 180$).**

### 3.2. Application II: Geometric program

Geometric programs (GP) have applications in a wide variety of areas including circuit design, optimal control, nonlinear network design and chemical equilibrium problems; see Boyd et al. (2007) for a tutorial on geometric programming and for a comprehensive list of applications.

The objective and constraints of GPs are described by *posynomial* functions of the form $f(z) = \sum_k c_k \prod_i z_i^{\alpha_{ik}}$, where $c_k > 0$ is a constant, $z_i > 0$ is a decision variable and $\alpha_{ik} \in \mathbb{R}$ is an exponent. The trick to solve these nonconvex programs is to use the transformation $z_i = e^{x_i}$ for all $i$ and replace $f(z)$ with $\log f(e^x)$ in the model. Such problems can be formulated as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \sum_k c_k e^{\boldsymbol{\alpha}_k^\intercal \boldsymbol{x}} \,\middle|\, \log f_j(e^x) \le 0, \forall j = 1, \dots, m \right\}, \tag{15}$$

where $f_j(z)$ is a posynomial function, and $\mathcal{X}$ is a polyhedron.

In a stochastic variant of (15), the exponents $\alpha_{ik}$ are random with distribution $g_{ik}(\xi_{ik}|\theta_{ik})$. Including all constraints in $\mathcal{X}$, we write the general form of the stochastic geometric program as

$$\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\xi|\theta} \left[ \sum_k c_k e^{\boldsymbol{\xi}_k^\intercal \boldsymbol{x}} \right], \tag{16}$$

where the expectation is taken with respect to the joint distribution of the random variables.

Under the assumption that the random exponents $\xi_{ik}$ are independent, the objective function of (16) decomposes into separable terms as $\mathbb{E}_{\xi|\theta}\left[\sum_k c_k e^{\boldsymbol{\xi}_k^\intercal \boldsymbol{x}}\right] = \sum_k c_k \prod_i \mathbb{E}_{\xi_{ik}|\theta_{ik}}\left[e^{\xi_{ik} x_i}\right]$. The unique property of such decomposition is that the term $\mathbb{E}_{\xi_{ik}|\theta_{ik}}\left[e^{\xi_{ik} x_i}\right]$ is equal to the *moment generating function* $M_{\xi_{ik}|\theta_{ik}}(x_i)$ of the distribution $g_{ik}(\xi_{ik}|\theta_{ik})$. As a result, the stochastic geometric problem (16) reduces to

$$\min_{\boldsymbol{x}\in\mathcal{X}} \sum_k c_k \prod_i M_{\xi_{ik}|\theta_{ik}}(x_i). \tag{17}$$

Assume now that the parameter $\theta_{ik}$ is random with prior $\pi(\theta_{ik})$. We next use the results of Section 2.1 to evaluate the Separate-EO and Joint-EO solution estimators for (17). We present the risk comparison under the linear loss (2).

Corollary 2. *Consider the stochastic geometric program* (17). *Assume that $\xi_{ik}$ has distribution $g_{ik}(\xi_{ik}|\theta_{ik})$ and its parameter $\theta_{ik}$ has prior $\pi(\theta_{ik})$. Further, assume that an observation $\bar{\xi}_{ik}$ is available. Then,*

*(i) $\hat{\boldsymbol{x}}^S(\bar{\xi}) \in \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} \sum_k c_k \prod_i M_{\xi_{ik}|\hat{\theta}^B_{ik}}(x_i)$, where $\hat{\theta}^B_{ik}$ is the Bayes estimator of $\theta_{ik}$ under the squared error loss.*

*(ii) $\hat{\boldsymbol{x}}^{J,L}(\bar{\xi}) \in \operatorname{argmin}_{\boldsymbol{x}\in\mathcal{X}} \sum_k c_k \prod_i M_{\xi_{ik}|\bar{\xi}_{ik}}(x_i)$, where $M_{\xi_{ik}|\bar{\xi}_{ik}}$ is the moment generating function of the posterior predictive distribution $h(\xi_{ik}|\bar{\xi}_{ik})$.*

For a numerical illustration, consider the stochastic geometric program (17) for the exponential-gamma conjugate pair with prior hyperparameters $\alpha_k$ and $\beta_k$. We assume that a set of observations $\bar{\xi}_k,\ k \in [K]$ for some $K \in \mathbb{Z}$, is available. According to Corollary 2, obtaining the Separate-EO solution estimator amounts to solving the convex program

$$z^S := \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \sum_{k=1}^{K} c_k \frac{\lambda_k(\bar{\xi}_k)}{\lambda_k(\bar{\xi}_k) - x_k} \right\}, \tag{18}$$

where $\lambda_k(\bar{\xi}_k) = \frac{\alpha_k+1}{\beta_k+\bar{\xi}_k}$. Here, we implicitly use the fact that the moment generating function of the exponential distribution is available in closed-form. However, this approach is not applicable for the Joint-EO method as the moment generating function of the posterior predictive distribution, Lomax, is known to be mathematically intractable. Therefore, we require to make use of sampling-based methods (such as SAA) to obtain the Joint-EO solution estimator. In particular, let $\tilde{\xi}_{rk}$ be a Lomax random variable[1] with parameters $\beta_k + \bar{\xi}_k$ and $\alpha_k + 1$, for $k \in [K]$ and $r \in [R]$, where $R$ denotes the sample size. Then applying the SAA method to (15) for the Lomax distribution, we can approximate the Joint-EO solution estimator with the solution of the following convex program:

$$z^{J,L}(R) := \min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \sum_{k=1}^{K} c_k \frac{1}{R} \sum_{r=1}^{R} e^{\tilde{\xi}_{rk} x_k} \right\}. \tag{19}$$

We now discuss an experimental setting. We define the feasible region $\mathcal{X}$ as the intersection of the standard simplex $\Delta_K := \{\boldsymbol{x} \in \mathbb{R}^K_+ :\ \sum_{k=1}^K x_k = 1\}$ with a polyhedron $\mathcal{P} := \{\boldsymbol{x} \in \mathbb{R}^K : A\boldsymbol{x} = \boldsymbol{b}\}$ where $A \in \mathbb{R}^{M\times K}$ and $\boldsymbol{b} \in \mathbb{R}^M$. Such linear constraints are common in geometric programs when the original posynomial function $f_j(z)$ contains a single summation, i.e., $f_j(z) = c\prod_i z_i^{\alpha_i}$. In this case, the above-mentioned geometric transformation yields $f_j(e^x) = ce^{\boldsymbol{\alpha}^\intercal \boldsymbol{x}}$, and hence the log constraint as presented in (15) becomes linear in $\boldsymbol{x}$, i.e, $\boldsymbol{\alpha}^\intercal\boldsymbol{x} \le -\log c$. We randomly generate the parameters according to the following rules:

$$\begin{aligned} c_k &= 1 & \alpha_k &\sim \text{Unif}(5,10) & \beta_k &\sim \text{Unif}(5,10) \\ A_{mk} &\sim \text{Unif}(1,10) & b_m &\sim \text{Unif}(10,50) & \bar{\xi}_k &= \frac{\alpha_k}{\beta_k}. \end{aligned}$$

In Table 1, we compare the Separate-EO and the Joint-EO solution estimators with $R = 100$, 1000 and 10000 for five randomly generated instances in dimension $K = 1000$ with $M = 10$ linear constraints. To compare the quality of the solutions obtained from the separate and approximate joint approach, we evaluate the solutions in the objective function of (19) for a new set of $R' = 10000$ independent samples generated from the Lomax distribution. We report these values in column

[1] Using the inverse transformation technique, we can generate such random variables as $\beta_k[(1-u)^{-1/(\alpha_k+1)} - 1]$, where $u$ is a uniform $[0,1]$ random variable

SEO for the Separate-EO and in columns JEO($R$) for the Joint-EO methods. The results give several interesting observations from different perspectives.

First, we observe that obtaining the JEO($R$) solution estimators becomes increasingly demanding with larger $R$ values in terms of the computational effort. For example, for moderate-sized instances considered here, it typically takes about 2 minutes to obtain a solution estimator when a sample size of 10000 is used. For these problem instances, we ran into memory issues with $R = 100000$ sample points. The SEO estimators, on the other hand, do not suffer from these issues since they are obtained as a solution to a deterministic convex program, which scales reasonably well with the instance size.

Second, because of the above issue, the objective values $z^{J,L}(R)$ for some instances of the Joint-EO method (instances 3, 4 and 5) exhibit a slow convergence. This causes a serious concern for determining a reliable stopping criterion for the SAA algorithms, as not only the objective values but also the optimal solutions are unstable for different sampling sizes. These obstacles for the computation of the Joint-EO solution estimators stem from the slow convergence result of the SAA method for certain problem structures. In particular, Birge (2016) argues that despite the asymptotic convergence of the SAA method with the growth of the sample size, the optimization problem is still prone to severe estimation errors when the number of uncertain elements (often linked to the problem scale) is large. In contrast, the Separate-EO solution estimators, despite being sub-optimal, enjoy a fast and stable solution process.

Third, unlike the pattern observed in the portfolio application, the approximate solution of the Joint-EO does not always give a better solution than that of the Separate-EO method. In particular, in instance 5, the SEO solution consistently has a better objective value than the JEO solution for different sizes. In instance 4, interestingly, the SEO solution has a better objective value than the JEO for sample sizes 100 and 1000; whereas for sample size 10000, the JEO solution dominates the SEO at the price of a larger computation time. This observation clearly shows that the computational limitations dictated by the problem structure and prior-likelihood distributions can change the dynamic between the performance of solution estimators obtained from the separate and joint approaches.

The applications presented in Sections 3.1 and 3.2 show a remarkable difference between the separate and joint methods when it comes to the accuracy versus computational efficiency trade-off. This suggests a closer study of these two methods when applied to various problem structures and probability distributions. We note here that if the separate and joint problems are solved approximately using sampling methods such as SAA, and both likelihood and posterior distributions are similarly convenient to sample from, the two models will have a similar computational complexity. In such a situation, the joint problem is clearly the preferred method. However, in this paper we

**Table 1 Computational results for randomly generated stochastic geometric programs with exponential-gamma pair ($K = 1000$, $M = 10$). Five problem instances (ins.) are solved, and the corresponding objective values (obj. val.) and computational times in seconds are recorded.**

| | SEO | | JEO (100) | | JEO (1000) | | JEO (10000) | |
|---|---|---|---|---|---|---|---|---|
| ins. | obj. val. | time | obj. val. | time | obj. val. | time | obj. val. | time |
| 1 | 1061.09 | 2.55 | 1009.50 | 3.57 | 1009.34 | 13.22 | 1009.16 | 131.27 |
| 2 | 5944.84 | 2.69 | 1016.91 | 3.79 | 1014.87 | 12.68 | 1014.39 | 100.33 |
| 3 | 76641.14 | 2.93 | 1313.29 | 3.65 | 1166.62 | 12.75 | 1064.67 | 95.55 |
| 4 | 2020.64 | 2.69 | 4891.68 | 3.73 | 2936.59 | 12.37 | 1547.22 | 89.58 |
| 5 | 1721.14 | 2.83 | 25154.60 | 3.58 | 4540.34 | 12.86 | 2121.23 | 110.28 |

investigate the exact solutions of these two problems which potentially leads to a difference in the computational effort, as illustrated in this section. In the remainder of this paper, we develop a theoretical foundation for such a comparison.

## 4. Risk comparison between the separate and joint estimators

Because of the trade-off observed in Section 3, two basic questions arise: (i) Under what conditions do the Separate-EO and Joint-EO methods lead to the same solution estimator? (ii) When the two estimators are different, is there a bound on the risk difference?

### 4.1. Sufficient conditions for the estimators to be the same

First, we present sufficient conditions under which the Separate-EO and the Joint-EO yield the same solution estimators for the linear loss (2).

PROPOSITION 4. *Assume that the objective function $\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi}, \boldsymbol{x})]$ of (1) is multilinear in $\boldsymbol{\theta}$, and that for any pair $(i, j) \in [n] \times [n]$, $i \neq j$ for which the product $\theta_i\theta_j$ appears in $\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi}, \boldsymbol{x})]$, we have that $\xi_i$ and $\xi_j$, as well as $\theta_i$ and $\theta_j$ are independent. Then, the set of Bayes solution estimators obtained from the Joint-EO scheme under the loss (2) is equal to the one obtained from the Separate-EO scheme, i.e., $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}}) = \hat{\boldsymbol{x}}^{S}(\bar{\boldsymbol{\xi}})$.*

Next, we present conditions for the quadratic loss (6).

PROPOSITION 5. *Assume that (1) has a unique optimal solution $\boldsymbol{x}^*(\boldsymbol{\theta})$ for any given $\boldsymbol{\theta}$. Assume also that each component of $\boldsymbol{x}^*(\boldsymbol{\theta})$ is multilinear in $\boldsymbol{\theta}$, and that for any pair $(i, j) \in [n] \times [n]$, $i \neq j$ for which the product $\theta_i\theta_j$ appears in a component, we have that $\xi_i$ and $\xi_j$, as well as $\theta_i$ and $\theta_j$ are independent. Then, the set of Bayes solution estimators obtained from the Joint-EO scheme under the loss (6) is equal to the one obtained from the Separate-EO scheme, i.e., $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}}) = \hat{\boldsymbol{x}}^{S}(\bar{\boldsymbol{\xi}})$.*

The proofs of these two propositions are given in Appendix B.

We give three examples of the portfolio selection model to illustrate how problem modeling and parameter distributions affect the solution estimators, leading to equal and/or different values.

In the first case all three solution estimators are the same.

EXAMPLE 1. For the classical Markowitz (mean-variance) formulation (Markowitz 1959)

$$\max_{\boldsymbol{x}\in\mathcal{X}} \ \boldsymbol{\mu}^\intercal \boldsymbol{x} - \frac{\tau}{2}\boldsymbol{x}^\intercal \Sigma \boldsymbol{x}, \tag{20}$$

assume $\mathcal{X} = \mathbb{R}^n$. This occurs when both long and short positions are allowed on all assets and a riskless asset is available.

Assume now that $\boldsymbol{\mu}$ is unknown and has a prior distribution $\pi(\boldsymbol{\mu})$, while $\Sigma$ is known and positive definite. The unconstrained problem (20) is convex and has a unique optimal solution $\boldsymbol{x}^*(\boldsymbol{\mu}) = \frac{1}{\tau}\Sigma^{-1}\boldsymbol{\mu}$. The conditions of Propositions 4 and 5 are satisfied. Therefore, all three solution estimators are equal to $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}}) = \hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}}) = \hat{\boldsymbol{x}}^{S}(\bar{\boldsymbol{\xi}}) = \frac{1}{\tau}\Sigma^{-1}\hat{\boldsymbol{\mu}}^{B}(\bar{\boldsymbol{\xi}})$ where $\hat{\boldsymbol{\mu}}^{B}(\bar{\boldsymbol{\xi}})$ is the Bayes estimator (posterior mean) of $\boldsymbol{\theta}$, which is a shrinkage vector under exponential distributions. ■

Next, we give an example where the solution estimators of the Separate-EO and the Joint-EO under the loss (2) are equal, but different from that of the Joint-EO under the loss (6).

EXAMPLE 2. Consider the variation of portfolio selection where we want to maximize expected return subject to bounding the risk of the portfolio. This problem can be modeled as (1), where $f(\boldsymbol{\xi}, \boldsymbol{x}) = \boldsymbol{\xi}^\intercal \boldsymbol{x}$, and $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^n | \boldsymbol{x}^\intercal \Sigma \boldsymbol{x} \leq r^2\}$. We obtain that

$$\max_{\boldsymbol{x}\in\mathcal{X}} \ \boldsymbol{\mu}^\intercal \boldsymbol{x}. \tag{21}$$

Assume now that $\boldsymbol{\mu}$ is unknown and has a prior distribution $\pi(\boldsymbol{\mu})$. Since the objective function of (21) is linear in $\boldsymbol{\mu}$, it follows from Proposition 4 that the solution estimators of the Separate-EO and the Joint-EO under the loss (2) are equal, i.e., $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}}) = \hat{\boldsymbol{x}}^{S}(\bar{\boldsymbol{\xi}}) = \frac{\Sigma^{-1}\hat{\boldsymbol{\mu}}^{B}(\bar{\boldsymbol{\xi}})}{||\Sigma^{-1/2}\hat{\boldsymbol{\mu}}^{B}(\bar{\boldsymbol{\xi}})||} r$. On the other hand, the unique optimal solution of (21) is obtained as $\boldsymbol{x}^*(\boldsymbol{\mu}) = \frac{\Sigma^{-1}\boldsymbol{\mu}}{||\Sigma^{-1/2}\boldsymbol{\mu}||} r$, which is not a linear function of $\boldsymbol{\mu}$. As a result, the conditions of Proposition 5 are not satisfied. This yields $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}}) = \mathbb{E}_{\mu|\bar{\xi}}\left[\frac{\Sigma^{-1}\boldsymbol{\mu}}{||\Sigma^{-1/2}\boldsymbol{\mu}||} r\right]$, a quantity that can be very difficult to compute depending on the distributions. ■

Finally, we illustrate the converse of Example 2, where the solution estimators of the Separate-EO and the Joint-EO under the loss (6) are equal, but different from that of the Joint-EO under the loss (2).

EXAMPLE 3. Consider the setting of Example 1, where the mean and covariance matrix of the asset returns are known, but the risk aversion factor $\tau$ is random with exponential distribution $\tau \sim \text{Exp}(\lambda)$. This problem can be modeled as (1), where $f(\boldsymbol{\tau}, \boldsymbol{x}) = \boldsymbol{\mu}^\intercal \boldsymbol{x} - \frac{\tau}{2}\boldsymbol{x}^\intercal \Sigma \boldsymbol{x}$, and $\mathcal{X} = \mathbb{R}^n$. We obtain that

$$\max_{\boldsymbol{x}\in\mathcal{X}} \ \boldsymbol{\mu}^\intercal \boldsymbol{x} - \frac{1}{2\lambda}\boldsymbol{x}^\intercal \Sigma \boldsymbol{x}. \tag{22}$$

Assume now that parameter $\lambda$ has a prior $\pi(\lambda)$. The unique optimal solution of the problem is $\boldsymbol{x}^*(\lambda) = \lambda\Sigma^{-1}\boldsymbol{\mu}$. Since $\boldsymbol{x}^*(\lambda)$ is linear in $\lambda$, it follows from Proposition 5 that the solution estimators of the Separate-EO and the Joint-EO under the loss (6) are equal, i.e., $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}}) = \hat{\boldsymbol{x}}^{S}(\bar{\boldsymbol{\xi}}) = \hat{\lambda}^B(\bar{\xi})\Sigma^{-1}\boldsymbol{\mu}$ where $\hat{\lambda}^B(\bar{\xi})$ is the Bayes estimator (posterior mean) of $\lambda$. On the other hand, since the objective function of (22) is not linear in $\lambda$, it does not satisfy the conditions of Proposition 4. In particular, $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}})$ is the maximizer of $\max_{\boldsymbol{x}\in\mathcal{X}}\ \boldsymbol{\mu}^\intercal\boldsymbol{x} - \mathbb{E}_{\lambda|\bar{\xi}}[\frac{1}{2\lambda}]\boldsymbol{x}^\intercal\Sigma\boldsymbol{x}$. This yields $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}}) = \frac{\Sigma^{-1}\boldsymbol{\mu}}{\mathbb{E}_{\lambda|\bar{\xi}}[\frac{1}{\lambda}]}$, which may be computable only numerically depending on the distributions. ■

### 4.2. The separate EO solution estimator can be arbitrarily poor

The next example shows that the Separate-EO solution can yield arbitrarily poor solutions under the linear loss (2).

Example 4. Assume that random variables $\xi_i$ for $i \in [n]$ are independent and follow a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ where $\mu_i$ is unknown and $\sigma_i^2$ is known. Assume also that the parameters $\mu_i$ for $i \in [n]$ follow a normal distribution $\mathcal{N}(\lambda_i, \delta_i^2)$ where $\lambda_i$ and $\delta_i^2$ are known. In this case, the posterior $\mu_i|\bar{\xi}_i$ has a normal distribution $\mathcal{N}(\eta_i, \zeta_i^2)$ where $\eta_i = \frac{\sigma_i^2}{\sigma_i^2+\delta_i^2}\lambda_i + \frac{\delta_i^2}{\sigma_i^2+\delta_i^2}\bar{\xi}_i$ and $\zeta_i^2 = \frac{\sigma_i^2\delta_i^2}{\sigma_i^2+\delta_i^2}$. Consider an instance of the stochastic program (1) where the objective function is $\mathbb{E}_{\xi|\mu}[f(\boldsymbol{\xi}, \boldsymbol{x})] = \sum_{i=1}^n \mu_i^2 x_i$ and $\mathcal{X}$ is an $n$-simplex, i.e.,

$$\max\left\{\sum_{i=1}^n \mu_i^2 x_i \,\middle|\, \sum_{i=1}^n x_i = 1, x_i \geq 0, \forall i \in [n]\right\}. \tag{23}$$

To obtain the Bayes solution estimator for the Joint-EO method under the linear loss (2), we solve (23) by computing the posterior expectation of its objective function, see (4). We obtain the objective function $\sum_{i=1}^n(\eta_i^2 + \zeta_i^2)x_i$ since $\mathbb{E}_{\mu|\bar{\xi}}[\mu_i^2] = \eta_i^2 + \zeta_i^2$. The optimal solution (Bayes solution estimator) is attained at $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}}) = \boldsymbol{e}^j$, where $\boldsymbol{e}^j$ is the $j^{\text{th}}$ unit vector and $j$ is the index of the maximum value among $\{\eta_i^2 + \zeta_i^2\}_{i\in[n]}$. In contrast, the minimizer (worst solution) of the above problem is attained at $\boldsymbol{x} = \boldsymbol{e}^k$, where $k$ is the index of the minimum value among $\{\eta_i^2 + \zeta_i^2\}_{i\in[n]}$. Now we calculate the estimator obtained from the Separate-EO method. The Bayes estimator of the unknown parameter $\boldsymbol{\mu}$ under the squared error loss is the posterior mean $\boldsymbol{\eta}$ and therefore the resulting objective function in (23) is $\sum_{i=1}^n \eta_i^2 x_i$. Using similar arguments as above, the optimal solution of this problem is $\hat{\boldsymbol{x}}^S(\bar{\boldsymbol{\xi}}) = \boldsymbol{e}^l$, where $l$ is the index of the maximum value among $\{\eta_i^2\}_{i\in[n]}$. In order to compare the quality of the estimators obtained from the Joint-EO and Separate-EO methods, we need to evaluate the objective value of $\boldsymbol{e}^l$ in the Joint-EO problem given above. For any values of the parameters $\boldsymbol{\sigma}$, $\boldsymbol{\delta}$ and data $\bar{\boldsymbol{\xi}}$ that satisfy $l = k$, the optimal solution of the Separate-EO method matches the worst solution in the Joint-EO problem. This shows that the Separate-EO estimator can be arbitrarily weak compared to the Joint-EO estimator.

For a numerical illustration, assume that $n=2$, $\bar{\xi}_1=2$, $\bar{\xi}_2=1$, $\lambda_1=\lambda_2=0$, $\sigma_1=\sigma_2=1$, $\delta_1=1$ and $\delta_2=3$. We compute $\eta_1=1$, $\eta_2=\frac{9}{10}$, $\zeta_1^2=\frac{1}{2}$ and $\zeta_2^2=\frac{9}{10}$. It follows that $\eta_1^2>\eta_2^2$ and $\eta_1^2+\zeta_1^2<\eta_2^2+\zeta_2^2$. This shows that the Bayes solution estimator for the Joint-EO is $\boldsymbol{x}^{J,L}(\bar{\boldsymbol{\xi}})=(0,1)$ and the solution estimator for the Separate-EO is $\boldsymbol{x}^{S}(\bar{\boldsymbol{\xi}})=(1,0)$. ■

A similar example is presented for the quadratic loss (6) in Appendix C. Even though these examples show that the Separate-EO method can yield arbitrarily poor solution estimators compared to the Joint-EO method, there are problem classes for which explicit bounds on the difference between these two estimators can be derived. We present a popular such class in the next section.

## 5. Piecewise linear functions

Piecewise linear functions encompass a variety of applications in optimization. They are used to model approximate nonlinear functions, constraints involving partial differential equations, discount policy in marketing, fixed-charge problems, etc; see Vielma et al. (2010). The stochastic variants of piecewise linear functions can be categorized into two major classes: one in which the stochasticity appears in the location of breakpoints (intercept), and one in which the stochasticity appears on the variable coefficients (slope). The former class includes stochastic newsvendor and median problems, while the latter includes rectifier activation functions in deep neural networks. In this section, we investigate examples of both of these classes using our Separate-EO and Joint-EO methods.

### 5.1. Piecewise linear functions with one breakpoint

Define $f(\xi,x):\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ as

$$f(\xi,x)=\begin{cases}\bar{f}(\xi,x), & \text{if } \xi\le x\\ \tilde{f}(\xi,x), & \text{if } \xi\ge x\end{cases}, \tag{24}$$

where $\bar{f}(\xi,x)=\bar{a}\xi+\bar{b}x+\bar{c}$, and $\tilde{f}(\xi,x)=\tilde{a}\xi+\tilde{b}x+\tilde{c}$. We assume that $f(\xi,x)$ is continuous over its domain including the breakpoint, which yields $\bar{f}(x,x)=\tilde{f}(x,x)$. In the above relation, $x$ is a decision variable, and $\xi$ is a random variable with distribution function $g(\xi|\theta)$, and $\theta$ is a random parameter with the prior distribution $\pi(\theta)$.

Such functions are commonly used to model inventory control and pricing problems. For instance, consider a single-stage newsvendor problem where the purchase cost per unit of a product is $c$ and the selling price per unit is $s$. The random demand $\xi$ follows a distribution $g(\xi|\theta)$ with a random parameter $\theta$ that has a prior $\pi(\theta)$. Let $x$ be the decision variable representing the order quantity. Then, the profit is $f(\xi,x)=s\min\{x,\xi\}-cx$ which is of the form (24) where $\bar{f}(\xi,x)=s\xi-cx$ and $\tilde{f}(\xi,x)=(s-c)x$. Despite being simple, the newsvendor problem encompasses a rich literature in optimization and is considered as one of the most fundamental structures in stochastic programming; see e.g., Chu et al. (2008), Liyanage and Shanthikumar (2005). Other areas of

application for piecewise linear functions with uncertain breakpoints emerge in modeling two-phase linear-linear processes such as latent growth behaviors Kohli and Harring (2013). These models are widely used to describe developmental processes in education and psychology.

Assume that the likelihood cumulative distribution function (cdf) $G$ satisfies $\lim_{t\to-\infty} tG(t|\theta)=0$, a property that holds for most of the standard probability distributions (such as normal, exponential and geometric distributions). Then, using an integration by part, the expectation of the function (24) is computed as

$$\begin{aligned}\mathbb{E}_{\xi|\theta}[f(\xi,x)] &= \left[\int_{z\le x}\bar{f}(z,x)g(z|\theta)dz+\int_{z\ge x}\tilde{f}(z,x)g(z|\theta)dz\right]\\ &= \tilde{a}\mathbb{E}_{\xi|\theta}[\xi]+\tilde{b}x+\tilde{c}+(\tilde{a}-\bar{a})\int_{z\le x}G(z|\theta)dz.\end{aligned} \tag{25}$$

It follows from (25) that the function $\mathbb{E}_{\xi|\theta}[f(\xi,x)]$, (i) is infinitely differentiable in $x$, (ii) is strictly convex if $\tilde{a}-\bar{a}>0$, strictly concave if $\tilde{a}-\bar{a}<0$ and linear if $\tilde{a}-\bar{a}=0$, and (iii) has a unique stationary point $x^*(\theta)=G^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\theta\right)$ when it is not linear, i.e., $\bar{a}-\tilde{a}\neq 0$.

Next, we consider the univariate unconstrained stochastic program

$$\max_{x\in\mathbb{R}}\mathbb{E}_{\xi|\theta}[f(\xi,x)], \tag{26}$$

where $f(\xi,x)$ is a piecewise linear objective function as defined in (24), and where $\bar{a}-\tilde{a}>0$ so that (26) is convex. Assume that the random variable $\xi$ follows distribution $g(\xi|\theta)$ and the parameter $\theta$ is random with a prior $\pi(\theta)$. Given an observation $\bar{\xi}$ from the distribution $g(\bar{\xi}|\theta)$, Corollary 3 presents the Separate-EO estimator $\hat{x}^S(\bar{\xi})$ and the Joint-EO estimators $\hat{x}^{J,L}(\bar{\xi})$ and $\hat{x}^{J,Q}(\bar{\xi})$ under the linear and quadratic loss. This result is obtained as a direct consequence of Proposition 2 and Corollary 1.

Corollary 3. *Consider the piecewise linear problem* (26)*, and assume that an observation* $\bar{\xi}$ *from the distribution* $g(\bar{\xi}|\theta)$ *is available. Then,*

*(i)* $\hat{x}^S(\bar{\xi})=G^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\hat{\theta}^B(\bar{\xi})\right)$*, where* $\hat{\theta}^B(\bar{\xi})$ *is the Bayes estimator of the unknown parameter* $\theta$ *under the squared error loss.*

*(ii)* $\hat{x}^{J,L}(\bar{\xi})=H^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\bar{\xi}\right)$*, where* $H(\xi|\bar{\xi})$ *is the cdf of the posterior predictive distribution of* $\xi$ *given the observation* $\bar{\xi}$*.*

*(iii)* $\hat{x}^{J,Q}(\bar{\xi})=\mathbb{E}_{\theta|\bar{\xi}}\left[G^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\theta\right)\right]$*, where the expectation is taken with respect to the posterior distribution* $\Pi(\theta|\bar{\xi})$*.*

It is clear from Corollary 3 that the solution estimators depend on the distribution of $\xi$ and its corresponding prior. As an example, we analyze the case with a normal-normal conjugate pair in detail.

PROPOSITION 6. *Consider the stochastic problem* (26)*. Assume that the likelihood distribution is normal with* $\xi \sim \mathcal{N}(\mu, \sigma^2)$ *and the prior distribution is normal with* $\mu \sim \mathcal{N}(\mu_0, \delta^2)$*. Assume further that the parameters* $\sigma^2$*,* $\mu_0$ *and* $\delta^2$ *are known, and an observation* $\bar{\xi}$ *is drawn from the likelihood distribution. Then, we have*

*(i)* $\hat{x}^S(\bar{\xi}) = \hat{x}^{J,Q}(\bar{\xi}) = [\rho\mu_0 + (1-\rho)\bar{\xi}] + \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\sigma$.

*(ii)* $\hat{x}^{J,L}(\bar{\xi}) = [\rho\mu_0 + (1-\rho)\bar{\xi}] + \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\sigma\sqrt{2-\rho}$.

*Here,* $\rho := \frac{\sigma^2}{\sigma^2+\delta^2}$ *and* $\Phi^{-1}(.)$ *is the inverse cdf of a standard normal random variable.*

*Proof:* (i) Due to Corollary 3(i), we have

$$\hat{x}^S(\bar{\xi}) = G^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\hat{\mu}^B(\bar{\xi})\right) = \hat{\mu}^B(\bar{\xi}) + \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\sigma,$$

where $G(\xi)$ is the cdf of a normal random variable with mean $\hat{\mu}^B(\bar{\xi}) := \rho\mu_0 + (1-\rho)\bar{\xi}$ and variance $\sigma^2$. Hence, we obtain $\hat{x}^S(\bar{\xi})$ as stated. We also observe that $\hat{x}^*(\mu)$ is linear in $\mu$. Therefore, by Proposition 5, we conclude that $\hat{x}^S(\bar{\xi}) = \hat{x}^{J,Q}(\bar{\xi})$.

(ii) Due to Corollary 3(ii), we have

$$\hat{x}^{J,L}(\bar{\xi}) = H^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\bar{\xi}\right) = \hat{\mu}^B(\bar{\xi}) + \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\sigma\sqrt{2-\rho},$$

where $H$ is the cdf of a normal random variable with mean $\hat{\mu}^B(\bar{\xi}) = \rho\mu_0 + (1-\rho)\bar{\xi}$ and variance $\sigma^2(2-\rho)$. Hence, the result follows. □

In the above proposition, the value of $\rho$ is based on $T = 1$ observation. The reader can verify that, in general, for $T$ observations, the proposition holds with $\rho := \frac{\sigma^2}{\sigma^2+T\delta^2}$. Similar adjustments can be made throughout this section. The computational complexity of obtaining the Separate-EO and Joint-EO estimators is similar as they both require an inverse normal cdf computation. We next provide bounds on the difference in their risk values. We will use the following result.

REMARK 1. Consider $d(\kappa) = \kappa\left(\sqrt{1+1/(1+\kappa^2)} - 1\right)$. Then,

$$(i)\ d(0) = 0,\ (ii)\ \lim_{\kappa\to\infty} d(\kappa) = 0,\ (iii)\ d^* := \max_{\kappa\geq 0} d(\kappa) \leq 0.22575.$$

PROPOSITION 7. *Let* $2\tilde{b} \geq \bar{a} - \tilde{a}$*. Under the assumptions of Proposition 6, we have*

$$\mathcal{R}^L\left(x^*(\mu), \hat{x}^S(\bar{\xi})\right) - \mathcal{R}^L\left(x^*(\mu), \hat{x}^{J,L}(\bar{\xi})\right) \leq \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\tilde{b}\sigma(\sqrt{2-\rho}-1) \leq \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\tilde{b}\delta d^*.$$

*Proof:* We first compute an upper bound on the difference in the loss values of the estimators. Define $F(x) := \mathbb{E}_{\xi|\mu}[f(\xi, x)]$ as the objective function of (26). Then, we have

$$\begin{aligned}
&\mathcal{L}^L(x^*(\theta), \hat{x}^S(\bar{\xi})) - \mathcal{L}^L(x^*(\theta), \hat{x}^{J,L}(\bar{\xi})) = \ F(\hat{x}^{J,L}(\bar{\xi})) - F(\hat{x}^S(\bar{\xi})) \\
&\leq F'(\hat{x}^S(\bar{\xi}))(\hat{x}^{J,L}(\bar{\xi}) - \hat{x}^S(\bar{\xi})) \leq \ \tilde{b}\Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\sigma(\sqrt{2-\rho}-1) \\
&= \tilde{b}\Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\delta d(\kappa) \leq \ \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\tilde{b}\delta d^*,
\end{aligned}$$

where the first equality follows from the definition of the linear loss (2), the first inequality is obtained from the first order Taylor expansion of the concave function $F(x)$ at point $\hat{x}^{J,L}(\bar{\xi})$ about $\hat{x}^{S}(\bar{\xi})$, the second inequality follows from (i) $\hat{x}^{J,L}(\bar{\xi}) - \hat{x}^{S}(\bar{\xi}) = \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\sigma(\sqrt{2-\rho}-1)$ because of Proposition 6, (ii) $\Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right) \geq 0$ because of the assumption $2\tilde{b} \geq \bar{a}-\tilde{a}$, and (iii) $F'(x) \leq \tilde{b}$ for all $x \in \mathbb{R}$ which is deduced from (25), the second equality holds due to the definition of $d(\kappa)$ with $\kappa := \sigma/\delta$ given in Remark 1 and $\rho = \frac{\sigma^2}{\sigma^2+\delta^2}$, and the last inequality follows from Remark 1(iii). Since the difference in the loss of the two estimators in the above expression is independent of both $\mu$ and $\bar{\xi}$, the risk difference obtained by taking the expectations $\mathbb{E}_{\mu}\mathbb{E}_{\bar{\xi}|\mu}$ as given in (3) remains unchanged, yielding the result. □

The following result is obtained similarly to that of Proposition 7.

PROPOSITION 8. *Let $2\tilde{b} \leq \bar{a}-\tilde{a}$. Under the assumptions of Proposition 6, we have*

$$\begin{aligned}\mathcal{R}^L\left(x^*(\mu), \hat{x}^S(\bar{\xi})\right) - \mathcal{R}^L\left(x^*(\mu), \hat{x}^{J,L}(\bar{\xi})\right) &\leq \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)(\tilde{b}+\tilde{a}-\bar{a})\sigma(\sqrt{2-\rho}-1) \\ &\leq \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)(\tilde{b}+\tilde{a}-\bar{a})\delta d^*.\end{aligned}$$

We also analyze the exponential-gamma and geometric-beta conjugate pairs for which the detailed derivations are provided in Appendix D.1. We present the algorithmic summary of these results in Table 2. This table clearly shows how different conjugate pairs affect computational complexity of the stochastic problem (26).

**Table 2 Summary of the computational effort required to obtain the Separate-EO and Joint-EO solution estimators for the univariate piecewise linear stochastic problem with different likelihood-prior pairs.**

| Likelihood | Prior | Separate-EO | Joint-EO (Linear Loss) | Joint-EO (Quadratic Loss) |
|---|---|---|---|---|
| Normal | Normal | inverse normal cdf | inverse normal cdf | inverse normal cdf |
| Exponential | Gamma | closed form | closed form | closed form |
| Geometric | Beta | closed form | need algorithm | need numerical integration |

### 5.2. Sum of piecewise linear functions with one breakpoint

In this section, we study the following extension of the piecewise linear model of Section 5.1

$$\max_{x\in\mathbb{R}^n} \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{\theta}}\left[\sum_{i=1}^{n} f_i(\xi_i, x)\right], \tag{27}$$

where $f_i(\xi_i, x)$ is a piecewise linear objective function of the form (24), i.e., $\bar{f}_i(\xi_i, x) = \bar{a}_i\xi_i + \bar{b}_i x + \bar{c}_i$ for $\xi_i \leq x$, and $\tilde{f}_i(\xi_i, x) = \tilde{a}_i\xi_i + \tilde{b}_i x + \tilde{c}_i$ for $\xi_i \geq x$. We assume that $\bar{a}_i - \tilde{a}_i > 0$ for each $i \in [n]$, so that (27) is convex. In this problem, $x$ is the single decision variable and $\xi_i$ is a random variable with

probability distribution $g_i(\xi_i|\theta_i)$ and the random parameter $\theta_i$ has prior distribution $\pi_i(\theta_i)$. In this setting, variables $\xi_i$ are not required to be independent. We further assume that an observation $\bar{\xi}_i$ drawn from $g_i(\bar{\xi}_i|\theta_i)$ for $i \in [n]$ is available.

Stochastic problems of the form (27) have applications in median and location-allocation problems in facility planning. For instance, consider the one-dimensional stochastic median problem, where the position $\xi_i$ of $n$ points on the line is random and each comes from a distribution $g_i(\xi_i|\theta_i)$ for $i \in [n]$. The goal is to find the location $x$ that minimizes the total distance from the random points, i.e., $\sum_{i=1}^{n} |\xi_i - x|$. This problem can be formulated as (27) where $f_i(\xi_i, x) = |\xi_i - x|$, $\bar{f}_i(\xi_i, x) = x - \xi_i$ and $\tilde{f}_i(\xi_i, x) = \xi_i - x$.

Using (25), the objective function of (27) can be computed as

$$\mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{\theta}}[\sum_{i=1}^{n} f_i(\xi_i, x)] = \sum_{i=1}^{n} \left[ \tilde{a}_i \mathbb{E}_{\xi_i|\theta_i}[\xi_i] + \tilde{b}_i x + \tilde{c}_i + (\tilde{a}_i - \bar{a}_i) \int_{z_i \le x} G_i(z_i|\theta_i) dz_i \right], \tag{28}$$

Since this is a concave function under the assumption $\bar{a}_i - \tilde{a}_i > 0$ for $i \in [n]$, its maximizer is obtained as the root of the equation

$$\sum_{i=1}^{n} (\tilde{a}_i - \bar{a}_i) G_i(x|\theta_i) + \sum_{i=1}^{n} \tilde{b}_i = 0. \tag{29}$$

We next give the Separate-EO estimator $\hat{x}^S(\bar{\xi})$ and the Joint-EO estimator $\hat{x}^{J,L}(\bar{\xi})$ via univariate root finding using (29). We do not present the Joint-EO estimator $\hat{x}^{J,Q}(\bar{\xi})$ since the explicit solution for (29) is not computable in general.

PROPOSITION 9. *Consider problem* (27). *Assume that, for each* $i \in [n]$, *the likelihood distribution has the pdf* $g_i(\xi_i|\theta_i)$ *and the prior distribution has the pdf* $\pi_i(\theta_i)$. *Assume further that an observation* $\bar{\xi}_i$ *is available for* $i \in [n]$. *Then,*

*(i)* $\hat{x}^S(\bar{\xi})$ *is the solution of*

$$\sum_{i=1}^{n} (\tilde{a}_i - \bar{a}_i) G_i(x|\hat{\theta}_i^B(\bar{\xi}_i)) + \sum_{i=1}^{n} \tilde{b}_i = 0,$$

*where* $\hat{\theta}_i^B(\bar{\xi}_i)$ *is the Bayes estimator of* $\theta_i$ *under the squared error loss, given observation* $\bar{\xi}_i$.

*(ii)* $\hat{x}^{J,L}(\bar{\xi})$ *is the solution of*

$$\sum_{i=1}^{n} (\tilde{a}_i - \bar{a}_i) H_i(x|\bar{\xi}_i) + \sum_{i=1}^{n} \tilde{b}_i = 0,$$

*where* $H_i(\xi|\bar{\xi}_i)$ *is the cdf of the posterior predictive distribution of* $\xi_i$ *given observation* $\bar{\xi}_i$.

Next, we give a generic method to compute an upper bound on the risk difference between the Separate-EO estimator $\hat{x}^S(\bar{\xi})$ and the Joint-EO estimator $\hat{x}^{J,L}(\bar{\xi})$.

PROPOSITION 10. *Consider the setting in Proposition 9. Assume that the likelihood $g_i(\xi_i|\theta_i)$ and the posterior predictive $h_i(\xi_i|\bar{\xi}_i)$ are positive at all points in the domain. Then,*

$$\begin{aligned}&\mathcal{R}^L\left(x^*(\boldsymbol{\theta}), \hat{x}^S(\bar{\xi})\right) - \mathcal{R}^L\left(x^*(\boldsymbol{\theta}), \hat{x}^{J,L}(\bar{\xi})\right)\\ &\leq K\mathbb{E}_{\bar{\boldsymbol{\xi}}}\big[\max\big\{\max_i\{\hat{x}_i^{J,L}(\bar{\xi}_i)\} - \min_i\{\hat{x}_i^S(\bar{\xi}_i)\}, \max_i\{\hat{x}_i^S(\bar{\xi}_i)\} - \min_i\{\hat{x}_i^{J,L}(\bar{\xi}_i)\}\big\}\big],\end{aligned}$$

*where the expectation $\mathbb{E}_{\bar{\boldsymbol{\xi}}}$ is taken with respect to the marginal distribution of $\bar{\xi}$, $K = \max\{|\sum_{i=1}^n \tilde{b}_i|, |\sum_{i=1}^n \tilde{b}_i + \tilde{a}_i - \bar{a}_i|\}$, $\hat{x}_i^S(\bar{\xi}_i)$ and $\hat{x}_i^{J,L}(\bar{\xi}_i)$ are the Separate-EO and Joint-EO estimators if there was only a single random variable $\xi_i$, as computed in Corollary 3.*

*Proof:* We will first obtain intervals which contain the solution estimators $\hat{x}^S(\bar{\xi})$ and $\hat{x}^{J,L}(\bar{\xi})$. Since $g_i(\xi_i|\theta_i)$ and $h_i(\xi_i|\bar{\xi}_i)$ are positive over the domain, their cdfs are strictly increasing. Therefore, we have that

$$\begin{aligned}G_i\left(x\middle|\hat{\theta}_i^B(\bar{\xi}_i)\right) &< \frac{\tilde{b}_i}{\bar{a}_i - \tilde{a}_i} \text{ for } x < G_i^{-1}\left(\frac{\tilde{b}_i}{\bar{a}_i - \tilde{a}_i}\middle|\hat{\theta}_i^B(\bar{\xi}_i)\right),\\ G_i\left(x\middle|\hat{\theta}_i^B(\bar{\xi}_i)\right) &> \frac{\tilde{b}_i}{\bar{a}_i - \tilde{a}_i} \text{ for } x > G_i^{-1}\left(\frac{\tilde{b}_i}{\bar{a}_i - \tilde{a}_i}\middle|\hat{\theta}_i^B(\bar{\xi}_i)\right),\end{aligned}$$

for each $i \in [n]$. Multiplying each of the above relations by $\bar{a}_i - \tilde{a}_i > 0$ and then summing them over $i$, we deduce that $\hat{x}^S(\bar{\xi}) \in \left[\min_i\left\{G_i^{-1}\left(\frac{\tilde{b}_i}{\bar{a}_i-\tilde{a}_i}\middle|\hat{\theta}_i^B(\bar{\xi}_i)\right)\right\}, \max_i\left\{G_i^{-1}\left(\frac{\tilde{b}_i}{\bar{a}_i-\tilde{a}_i}\middle|\hat{\theta}_i^B(\bar{\xi}_i)\right)\right\}\right]$. A similar argument shows that $\hat{x}^{J,L}(\bar{\xi}) \in \left[\min_i\left\{H_i^{-1}\left(\frac{\tilde{b}_i}{\bar{a}_i-\tilde{a}_i}\middle|\bar{\xi}_i\right)\right\}, \max_i\left\{H_i^{-1}\left(\frac{\tilde{b}_i}{\bar{a}_i-\tilde{a}_i}\middle|\bar{\xi}_i\right)\right\}\right]$.

We now give an upper bound on the loss values of these estimators. Let $F(x) = \mathbb{E}_{\boldsymbol{\xi}|\boldsymbol{\theta}}\left[\sum_{i=1}^n f_i(\xi_i, x)\right]$ as the objective function of (27). Since $F$ is concave, we have

$$\begin{aligned}&\mathcal{L}^L(x^*(\theta), \hat{x}^S(\bar{\xi})) - \mathcal{L}^L(x^*(\theta), \hat{x}^{J,L}(\bar{\xi})) = \ F(\hat{x}^{J,L}(\bar{\xi})) - F(\hat{x}^S(\bar{\xi}))\\ &\leq F'(\hat{x}^S(\bar{\xi}))(\hat{x}^{J,L}(\bar{\xi}) - \hat{x}^S(\bar{\xi})) \leq \ |F'(\hat{x}^S(\bar{\xi}))||\hat{x}^{J,L}(\bar{\xi}) - \hat{x}^S(\bar{\xi})|\\ &\leq K\max\big\{\max_i\{\hat{x}_i^{J,L}(\bar{\xi}_i)\} - \min_i\{\hat{x}_i^S(\bar{\xi}_i)\}, \max_i\{\hat{x}_i^S(\bar{\xi}_i)\} - \min_i\{\hat{x}_i^{J,L}(\bar{\xi}_i)\}\big\},\end{aligned}$$

where the last inequality holds because (i) $|F'(\hat{x}^S)| \leq K$ as $F'(\hat{x}^S) \in [\sum_{i=1}^n \tilde{b}_i + \tilde{a}_i - \bar{a}_i, \sum_{i=1}^n \tilde{b}_i]$ due to (28), and (ii) the previously derived intervals for $\hat{x}^S(\bar{\xi})$ and $\hat{x}^{J,L}(\bar{\xi})$, and the fact that: if $w \in [w_1, w_2]$ and $v \in [v_1, v_2]$, then $|w - v| \leq \max\{w_2 - v_1, v_2 - w_1\}$. To compute an upper bound on the risk difference, we take the expectation $\mathbb{E}_\theta\mathbb{E}_{\bar{\xi}|\theta}$ or equivalently $\mathbb{E}_{\bar{\xi}}\mathbb{E}_{\theta|\bar{\xi}}$ of both sides of the above expression. Since this expression is independent of $\theta$ the desired result is obtained by only taking the expectation $\mathbb{E}_{\bar{\xi}}$. □

The result of Proposition 10 requires an additional expectation with respect to the marginal distribution of $\bar{\xi}$, over the maximum of random variables. As a consequence, the bound given in this proposition can be hard to compute in closed-form for general classes of the stochastic program (27). However, it can be computed explicitly for certain objective functions and probability distributions.

To illustrate the derivation technique, we next consider the one-dimensional stochastic median problem where $f_i(\xi_i, x) = |\xi_i - x|$ for $i \in [n]$, and compute the bound on the risk of the Separate-EO and Joint-EO solution estimators for normal-normal conjugate pair. An analysis with exponential-gamma pair is provided in Appendix D.2.

PROPOSITION 11. *Consider the one-dimensional stochastic median problem. Assume that, for each $i \in [n]$, the likelihood distribution is normal with $\xi_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and the prior distribution is normal with $\mu_i \sim \mathcal{N}(\mu_i^0, \delta_i^2)$. Assume further that the parameters $\sigma_i^2$, $\mu_i^0$, $\delta_i^2$ are known, and a realization of locations $\bar{\xi}_i$ is observed. Then, we have*

$$\mathcal{R}^L\left(x^*(\boldsymbol{\mu}), \hat{x}^S(\bar{\xi})\right) - \mathcal{R}^L\left(x^*(\boldsymbol{\mu}), \hat{x}^{J,L}(\bar{\xi})\right) \leq n\left(\max_i\{\mu_i^0\} - \min_i\{\mu_i^0\} + 2\sqrt{\frac{n-1}{n}\sum_{i=1}^n \frac{\delta_i^4}{\sigma_i^2+\delta_i^2}}\right).$$

*Proof:* Proposition 6 implies that $\nu_i := \hat{x}_i^S(\bar{\xi}_i) = \hat{x}_i^{J,L}(\bar{\xi}_i) = \rho_i\mu_i^0 + (1-\rho_i)\bar{\xi}_i$ where $\rho_i := \frac{\sigma_i^2}{\sigma_i^2+\delta_i^2}$ for $i \in [n]$. This result follows from the facts that $\frac{\tilde{b}_i}{\bar{a}_i - \tilde{a}_i} = \frac{1}{2}$ for the median objective function, and therefore $\Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right) = 0$. Next, we use Proposition 10 to obtain

$$\mathcal{R}^L\left(x^*(\boldsymbol{\mu}), \hat{x}^S(\bar{\xi})\right) - \mathcal{R}^L\left(x^*(\boldsymbol{\mu}), \hat{x}^{J,L}(\bar{\xi})\right) \leq n\mathbb{E}_{\bar{\boldsymbol{\xi}}}[\max_i\{\nu_i\} - \min_i\{\nu_i\}]. \tag{30}$$

Now, we compute an upper bound on the right-hand-side using a well-known result in probability theory to bound max/min expectations. In particular, for random variables $z_1, \ldots, z_n$, we have $\mathbb{E}[\max_i\{z_i\}] \leq \max_i\{\mathbb{E}[z_i]\} + \sqrt{\frac{n-1}{n}\sum_{i=1}^n \mathrm{Var}(z_i)}$ and $\mathbb{E}[\min_i\{z_i\}] \geq \min_i\{\mathbb{E}[z_i]\} - \sqrt{\frac{n-1}{n}\sum_{i=1}^n \mathrm{Var}(z_i)}$; see Aven (1985). Since $\bar{\xi}_i \sim \mathcal{N}(\mu_i^0, \sigma_i^2 + \delta_i^2)$ for each $i \in [n]$, we have $\mathbb{E}[\nu_i] = \mu_i^0$ and $\mathrm{Var}(\nu_i) = \frac{\delta_i^4}{\sigma_i^2+\delta_i^2}$. Using these relations in the above bounding inequalities, we obtain

$$\mathbb{E}[\max_i\{\nu_i\}] \leq \max_i\{\mu_i^0\} + \sqrt{\frac{n-1}{n}\sum_{i=1}^n \frac{\delta_i^4}{\sigma_i^2+\delta_i^2}} \quad \text{and} \quad \mathbb{E}[\min_i\{\nu_i\}] \geq \min_i\{\mu_i^0\} - \sqrt{\frac{n-1}{n}\sum_{i=1}^n \frac{\delta_i^4}{\sigma_i^2+\delta_i^2}}.$$

Plugging in these bounds into (30) gives the desired conclusion. □

### 5.3. Piecewise linear functions with random slope

In this section we consider a stochastic piecewise linear model that includes randomness in the coefficients of the decision variables. This class of problems models the *rectified linear unit* (ReLU) activation function often used in neural networks; see Goodfellow et al. (2016). Unlike for the class studied in Sections 5.1 and 5.2, an explicit derivation of Separate-EO and Joint-EO solution estimators seems out of reach in this case. Therefore we illustrate the difference in the quality of solutions obtained from the separate and joint methods through a computational experiment in a streamlined neural network model with a single layer.

Suppose that there are $n$ features denoted by random variables $\xi_1, \ldots, \xi_n$ that define an output $O$ through the following procedure

$$O = \max\big\{0, \sum_{j=1}^{n} \xi_j w_j + w_{n+1} + \epsilon\big\}, \tag{31}$$

where $w_1, \ldots, w_n, w_{n+1}$ are the true weights of the random features and $\epsilon$ is a random error. The goal is to model this problem using a single-layer neural network and to evaluate the accuracy of the Separate-EO and Joint-EO methods in estimating the weights.

To this end, we consider a setting where the likelihood and prior distributions of $\xi_i$ are independent normal distributions with known variance and prior mean, that is, $\xi_j|\mu_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ and $\mu_j \sim \mathcal{N}(\mu_{0j}, \delta_{0i})$, $j = 1, \ldots, n$. Further, we assume that $T$ observations drawn from the likelihood are available as $\bar{\boldsymbol{\xi}}^1, \ldots, \bar{\boldsymbol{\xi}}^T$. For such a conjugate pair the posterior and predictive distributions are readily available.

The corresponding single-layer neural network model with $n$ input nodes and a single output node is represented as

$$O = \max\big\{0, \sum_{j=1}^{n} \xi_j x_j + x_{n+1}\big\}, \tag{32}$$

where $x_1, \ldots, x_n, x_{n+1}$ are decision variables. To determine these variables, we need to solve the problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^{n+1}} \mathbb{E}_{\xi|\mu} \Big| O - \max\big\{0, \sum_{j=1}^{n} \xi_j x_j + x_{n+1}\big\} \Big|. \tag{33}$$

The Separate-EO estimator can be obtained as an optimal solution of

$$\min_{\boldsymbol{x} \in \mathbb{R}^{n+1}} \mathbb{E}_{\xi|\mu^B(\bar{\xi})} \Big| O - \max\big\{0, \sum_{j=1}^{n} \xi_j x_j + x_{n+1}\big\} \Big|, \tag{34}$$

where $\mu^B(\bar{\xi})$ is the Bayes estimator of $\mu$ under the quadratic loss, whereas the Joint-EO estimator under the linear loss can computed as an optimal solution of

$$\min_{\boldsymbol{x} \in \mathbb{R}^{n+1}} \mathbb{E}_{\xi|\bar{\xi}} \Big| O - \max\big\{0, \sum_{j=1}^{n} \xi_j x_j + x_{n+1}\big\} \Big|. \tag{35}$$

Since solving problems (34) and (35) directly does not seem possible, we adopt a sampling approach as common in machine learning. In particular, assuming that $R$ sample points are available from the posterior and predictive distributions, we define an approximate problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^{n+1}} \frac{1}{R} \sum_{r=1}^{R} \Big| O^r - \max\big\{0, \sum_{j=1}^{n} \xi_j^r x_j + x_{n+1}\big\} \Big|, \tag{36}$$

which minimizes the average prediction error over the sample space. We denote the optimal solutions of problem (36) as $\hat{\boldsymbol{x}}^S$ and $\hat{\boldsymbol{x}}^{J,L}$ when sample points are obtained from the posterior and predictive distributions, respectively.

To conduct our experiments, we draw sample points of input features from the posterior and predictive distributions and compute the sample output according to (31) using true weights. We first randomly generate the parameters according to the following rules:

$$w_j, w_{n+1} \sim \text{Unif}(-1,1),\ \mu_{0j} \sim \text{Unif}(-1,1),\ \delta_{0j} \sim \text{Unif}(0,0.1),\ \sigma_j \sim \text{Unif}(0.5,\bar{\sigma}),\ j=1,\dots,n.$$

We also set the distribution of the error term as $\epsilon \sim \mathcal{N}(0, 0.1^2)$.

We solve problem (36) with $R = 1000$ as a mixed-integer linear program after linearizing the objective. This requires assuming an upper bound on the quantity $\sum_{j=1}^{n} \xi_j x_j + x_{n+1}$, which we set to $10^5$. Once the optimal weights $\hat{\boldsymbol{x}}^S$ and $\hat{\boldsymbol{x}}^{J,L}$ are calculated, we compute the test error on a new (test) dataset with size $R$ again. The average test error results of 100 macro-replications are reported in Table 3. We observe that the test errors with the Joint-EO approach are significantly lower than those of the Separate-EO approach. For the same problem and sample size, the difference becomes even more drastic when the upper bound parameter of the likelihood standard deviation $\bar{\sigma}$ becomes larger. The extension of such computational studies performed on a deeper neural net with more complex structures while employing different simulation techniques is left as a direction of future research.

**Table 3 Test error results in the deep learning application with different problem sizes (here, SEO and JEO represent Separate-EO and Joint-EO, respectively).**

| | | $\bar{\sigma}=1.0$ | | $\bar{\sigma}=1.5$ | | $\bar{\sigma}=2.0$ | | $\bar{\sigma}=2.5$ | |
|---|---|---|---|---|---|---|---|---|---|
| $n$ | $T$ | SEO | JEO | SEO | JEO | SEO | JEO | SEO | JEO |
| 10 | 100 | 0.876 | 0.114 | 34.242 | 0.097 | 1.356 | 0.123 | 2.911 | 0.122 |
| 20 | 200 | 1.454 | 0.110 | 2.011 | 0.111 | 2.371 | 0.108 | 5.469 | 0.127 |
| 30 | 300 | 3.455 | 0.105 | 5.002 | 0.098 | 3.624 | 0.124 | 4.779 | 0.179 |
| 40 | 400 | 1.481 | 0.099 | 3.696 | 0.105 | 10.625 | 0.111 | 4.746 | 0.239 |
| 50 | 500 | 2.557 | 0.093 | 5.322 | 0.103 | 16.695 | 0.110 | 8.038 | 0.390 |

## 6. Conclusion

In the presence of uncertainty, data is used to estimate the unknown elements of stochastic programs. Several techniques can be employed. Under a Bayesian framework, two of the most common estimation rules solve the estimation and optimization problems either separately or jointly. In this paper we explore the analytical and computational trade-offs between the quality and complexity of these schemes in parametric stochastic programs. We use risk as an averaging measure for the quality of the solutions based on two optimization criteria: the gap between the objective values, and the distance between the solutions. We show conditions under which the two schemes yield equal solutions, and give examples when the risk difference between the solutions of the two schemes can be very large. We further study several classes of nonlinear stochastic programs with a piecewise linear structure, while presenting computational experiments on various applications.

## Appendix A: Omitted proofs from Section 2

*Proof of Proposition 1* We write that

$$\mathcal{R}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})) = \mathbb{E}_{\theta}\mathbb{E}_{\bar{\xi}|\theta}\left[\mathcal{L}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))\right] = \mathbb{E}_{\bar{\xi}}\mathbb{E}_{\theta|\bar{\xi}}\left[\mathcal{L}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))\right],$$

where the first equality follows from the definition of the linear risk, and the second equality follows from exchanging the order of sequential expectations. According to the definition of Bayes solution estimator, we seek among all $\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}) \in \mathcal{X}$ an estimator that minimizes the risk $\mathcal{R}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))$. First, we claim that any minimizer $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}})$ of $\mathbb{E}_{\theta|\bar{\xi}}\left[\mathcal{L}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))\right]$ is also a minimizer of the risk $\mathcal{R}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))$. To prove this claim, consider any estimator $\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}) \neq \hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}})$. It follows from the assumption that $\mathbb{E}_{\theta|\bar{\xi}}\left[\mathcal{L}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}}))\right] \leq \mathbb{E}_{\theta|\bar{\xi}}\left[\mathcal{L}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))\right]$. Taking the expectation $\mathbb{E}_{\bar{\xi}}[.]$ with respect to the marginal distribution of $\bar{\boldsymbol{\xi}}$ from both sides, we obtain that $\mathbb{E}_{\bar{\xi}}\mathbb{E}_{\theta|\bar{\xi}}\left[\mathcal{L}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}}))\right] \leq \mathbb{E}_{\bar{\xi}}\mathbb{E}_{\theta|\bar{\xi}}\left[\mathcal{L}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))\right]$. Hence, the claim follows from the chain relation in the first line. As a result, a Bayes solution estimator is a minimizer $\hat{\boldsymbol{x}}^{J,L}(\bar{\boldsymbol{\xi}})$ of $\mathbb{E}_{\theta|\bar{\xi}}\left[\mathcal{L}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))\right] = \mathbb{E}_{\theta|\bar{\xi}}\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi},\boldsymbol{x}^*(\boldsymbol{\theta}))] - \mathbb{E}_{\theta|\bar{\xi}}\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi},\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))]$. Recall that $\boldsymbol{x}^*(\boldsymbol{\theta})$ is independent of $\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})$, i.e., the term $\mathbb{E}_{\theta|\bar{\xi}}\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi},\boldsymbol{x}^*(\boldsymbol{\theta}))]$ is fixed regardless of the value of $\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})$. Therefore, any maximizer of $\mathbb{E}_{\theta|\bar{\xi}}\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi},\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))]$ is also a minimizer of $\mathbb{E}_{\theta|\bar{\xi}}\left[\mathcal{L}^L(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))\right]$. □

*Proof of Proposition 2* Using the result of Proposition 1, it suffices to show that the objective function of (4) matches that of (5). Using Fubini's theorem of integration, the assumption that $\mathbb{E}_{\xi|\bar{\xi}}[|f(\boldsymbol{\xi},\boldsymbol{x})|]$ is finite for any $\boldsymbol{x} \in \mathcal{X}$ allows for the interchange of the integration order. We have that

$$\begin{aligned}\mathbb{E}_{\theta|\bar{\xi}}\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi},\boldsymbol{x})] &= \int_{\boldsymbol{\theta}}\int_{\boldsymbol{\xi}} f(\boldsymbol{\xi},\boldsymbol{x})g(\boldsymbol{\xi}|\boldsymbol{\theta})\Pi(\boldsymbol{\theta}|\bar{\boldsymbol{\xi}})d\boldsymbol{\xi}d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\xi}} f(\boldsymbol{\xi},\boldsymbol{x})\left(\int_{\boldsymbol{\theta}} g(\boldsymbol{\xi}|\boldsymbol{\theta})\Pi(\boldsymbol{\theta}|\bar{\boldsymbol{\xi}})d\boldsymbol{\theta}\right)d\boldsymbol{\xi} = \int_{\boldsymbol{\xi}} f(\boldsymbol{\xi},\boldsymbol{x})h(\boldsymbol{\xi}|\bar{\boldsymbol{\xi}})d\boldsymbol{\xi} = \mathbb{E}_{\xi|\bar{\xi}}[f(\boldsymbol{\xi},\boldsymbol{x})],\end{aligned}$$

where the second equality is obtained by the interchange of integration order, and the third equality follows from the definition of posterior predictive distribution in Definition 1. □

*Proof of Proposition 3* We have that $\mathcal{R}^Q(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})) = \mathbb{E}_{\theta}\mathbb{E}_{\bar{\xi}|\theta}[\mathcal{L}^Q(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))] = \mathbb{E}_{\bar{\xi}}\mathbb{E}_{\theta|\bar{\xi}}[\mathcal{L}^Q(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))]$, where the first equality holds because of (7), and the second equality follows from changing the order of sequential expectations. According to the definition of Bayes estimator, we seek among all $\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}) \in \mathcal{X}$ an estimator that minimizes the risk $\mathcal{R}^Q(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))$. We claim that any minimizer $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}})$ of $\min_{\boldsymbol{x}\in\mathcal{X}}\mathbb{E}_{\theta|\bar{\xi}}[\mathcal{L}^Q(\boldsymbol{x}^*(\boldsymbol{\theta}),\boldsymbol{x})]$ is also a minimizer of the risk $\mathcal{R}^Q(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))$. To prove this claim, consider any estimator $\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}) \neq \hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}})$. It follows from the assumption that $\mathbb{E}_{\theta|\bar{\xi}}[\mathcal{L}^Q(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}}))] \leq \mathbb{E}_{\theta|\bar{\xi}}[\mathcal{L}^Q(\boldsymbol{x}^*(\boldsymbol{\theta}),\hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}}))]$. Taking the expectation $\mathbb{E}_{\bar{\xi}}[.]$ with respect to the marginal distribution of $\bar{\boldsymbol{\xi}}$ from both sides we obtain the desired result due to the chain relation in the first line. □

*Proof of Corollary 1* Since $\boldsymbol{x}^*(\boldsymbol{\theta})$ is unique, we obtain that $\mathcal{D}^2\big(\boldsymbol{x}^*(\boldsymbol{\theta}), \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})\big) = ||\boldsymbol{x}^*(\boldsymbol{\theta}) - \hat{\boldsymbol{x}}(\bar{\boldsymbol{\xi}})||^2$. To obtain the Bayes solution estimator, we need to find a minimizer of (8) which reduces to $\min_{\boldsymbol{x}\in\mathcal{X}} \mathbb{E}_{\theta|\bar{\xi}}[||\boldsymbol{x}^*(\boldsymbol{\theta}) - \boldsymbol{x}||^2]$. Define $\boldsymbol{w} = \mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\theta})]$. We write that

$$\begin{aligned}
\mathbb{E}_{\theta|\bar{\xi}}\left[||\boldsymbol{x}^*(\boldsymbol{\theta}) - \boldsymbol{x}||^2\right] &= \mathbb{E}_{\theta|\bar{\xi}}\left[||(\boldsymbol{x}^*(\boldsymbol{\theta}) - \boldsymbol{w}) + (\boldsymbol{w} - \boldsymbol{x})||^2\right] \\
&= \mathbb{E}_{\theta|\bar{\xi}}\left[||\boldsymbol{x}^*(\boldsymbol{\theta}) - \boldsymbol{w}||^2\right] + \mathbb{E}_{\theta|\bar{\xi}}\left[||\boldsymbol{w} - \boldsymbol{x}||^2\right] + 2\mathbb{E}_{\theta|\bar{\xi}}\left[(\boldsymbol{x}^*(\boldsymbol{\theta}) - \boldsymbol{w})^\intercal(\boldsymbol{w} - \boldsymbol{x})\right] \\
&= \mathbb{E}_{\theta|\bar{\xi}}\left[||\boldsymbol{x}^*(\boldsymbol{\theta}) - \boldsymbol{w}||^2\right] + ||\boldsymbol{w} - \boldsymbol{x}||^2,
\end{aligned}$$

where the first equality is obtained by adding and subtracting $\boldsymbol{w}$, the second equality follows from decomposition of the norm vector, and the last equality holds because $\mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\theta}) - \boldsymbol{w}] = \boldsymbol{0}$ by definition, and because $||\boldsymbol{w} - \boldsymbol{x}||^2$ does not depend on $\boldsymbol{\theta}$. Note in the last relation that the first term $\mathbb{E}_{\theta|\bar{\xi}}[||\boldsymbol{x}^*(\boldsymbol{\theta}) - \boldsymbol{w}||^2]$ does not contain $\boldsymbol{x}$. This gives the desired relation (9). For the next result, when $\mathcal{X}$ is convex, any convex combination of $\boldsymbol{x}^*(\boldsymbol{\theta})$ belongs to $\mathcal{X}$ as $\boldsymbol{x}^*(\boldsymbol{\theta}) \in \mathcal{X}$. We conclude that $\mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\theta})] \in \mathcal{X}$, and therefore the minimizer of (9) is attained at $\mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\theta})]$. □

## Appendix B: Omitted proofs from Section 4

*Proof of Proposition 4* It follows from the assumptions that $\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi}, \boldsymbol{x})]$ can be written as $\sum_{k=1}^K h_k(\boldsymbol{x}) g_k(\boldsymbol{\theta})$ where $h_k(\boldsymbol{x}) : \mathbb{R}^m \to \mathbb{R}$ and $g_k(\boldsymbol{\theta}) = \prod_{i\in I_k} \theta_i$ for some $I_k \subseteq [n]$. For this function, the assumption also implies that for any $k \in \{1, \ldots, K\}$, all variables $\xi_l$ for $l \in I_k$ are independent, and so are all variables $\theta_l$ for $l \in I_k$. We compute the objective of the Joint-EO method as given in (4)

$$\mathbb{E}_{\theta|\bar{\xi}}\mathbb{E}_{\xi|\theta}[f(\boldsymbol{\xi}, \boldsymbol{x})] = \sum_{k=1}^K h_k(\boldsymbol{x}) \mathbb{E}_{\theta|\bar{\xi}}[g_k(\boldsymbol{\theta})] = \sum_{k=1}^K h_k(\boldsymbol{x}) \prod_{i\in I_k} \mathbb{E}_{\theta_i|\bar{\xi}_i}[\theta_i], \tag{37}$$

where the first equality follows from linearity of the expectation operator and the second equality holds because random variables and parameters are independent. As discussed before, the Bayes estimator for a parameter under the squared error loss is the posterior mean. Therefore for the Separate-EO method, each $\theta_i$ is replaced by its posterior mean $\mathbb{E}_{\theta_i|\bar{\xi}_i}[\theta_i]$ in (1). The resulting objective function is (37). This shows that the objective of the Joint-EO and Separate-EO methods are equal. Since both have the same constraint set $\mathcal{X}$, their optimal solutions are the same. □

*Proof of Proposition 5* Under the assumption that $\boldsymbol{x}^*(\boldsymbol{\theta})$ is unique for any given $\boldsymbol{\theta}$, the Separate-EO solution estimator can be expressed as $\hat{\boldsymbol{x}}^S(\bar{\boldsymbol{\xi}}) = \boldsymbol{x}^*(\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{\xi}}))$ where $\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{\xi}})$ is the Bayes estimator of the unknown parameter $\boldsymbol{\theta}$ under the squared error loss. It follows from the discussion in Section 2.1 that $\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{\xi}}) = \mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{\theta}]$. Therefore, the Separate-EO method yields $\hat{\boldsymbol{x}}^S(\bar{\boldsymbol{\xi}}) = \boldsymbol{x}^*(\mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{\theta}])$. We obtain that $\mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\theta})] = \boldsymbol{x}^*(\mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{\theta}])$ because of the linearity of the expectation operator and because of the independence of variables appearing in the products. As a result, $\mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\theta})] \in \mathcal{X}$. It follows from (9) in Corollary 1 that the Joint-EO method yields $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}}) = \mathbb{E}_{\theta|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\theta})]$. □

## Appendix C: Omitted example from Section 4

EXAMPLE 5. Assume the setting of Example 4. Consider an instance of stochastic program (1) where the objective function is $\mathbb{E}_{\xi|\mu}[f(\boldsymbol{\xi},\boldsymbol{x})]=\sum_{i=1}^n \mu_i x_i$ and $\mathcal{X}$ is a unit-ball in $\mathbb{R}^n$, i.e.,

$$\max\left\{\sum_{i=1}^n \mu_i x_i \;\middle|\; \sum_{i=1}^n x_i^2 \le 1\right\}. \tag{38}$$

For any $\boldsymbol{\mu}\in\mathbb{R}^n\setminus\{0\}$, the unique optimal solution of the problem is $\boldsymbol{x}^*(\boldsymbol{\mu})=\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}$. As a result, the Separate-EO method yields the solution estimator $\boldsymbol{x}^S(\bar{\boldsymbol{\xi}})=\frac{\boldsymbol{\eta}}{||\boldsymbol{\eta}||}$. Now we calculate the estimator obtained from the Joint-EO method under the quadratic loss. Since the optimal solution of (38) is unique and its feasible region is convex, it follows from Corollary 1 that $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}})=\mathbb{E}_{\mu|\bar{\xi}}[\boldsymbol{x}^*(\boldsymbol{\mu})]=\mathbb{E}_{\mu|\bar{\xi}}\left[\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}\right]$. In particular, the Joint-EO solution is a convex combination of normalized vectors $\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}$ for all possible values of $\boldsymbol{\mu}$ taken according to the posterior distribution of $\boldsymbol{\mu}|\bar{\boldsymbol{\xi}}$. When this distribution is non-degenerate (can assume multiple distinct values), the expected vector $\hat{\boldsymbol{x}}^{J,Q}(\bar{\boldsymbol{\xi}})$ belongs to the interior of the unit-ball. On the other hand, the Separate-EO solution $\boldsymbol{x}^S(\bar{\boldsymbol{\xi}})$ is always on the boundary of the unit-ball. We conclude that the two solution estimators can never be equal. Moreover, they can achieve a maximum distance in the unit-ball. For instance, assume that the parameters $\boldsymbol{\lambda}$, $\boldsymbol{\sigma}$, $\boldsymbol{\delta}$ and the observation $\bar{\boldsymbol{\xi}}$ are such that $|\eta_i|=\left|\frac{\sigma_i^2}{\sigma_i^2+\delta_i^2}\lambda_i+\frac{\delta_i^2}{\sigma_i^2+\delta_i^2}\bar{\xi}_i\right|\le\epsilon$ for all $i\in[n]$ and a sufficiently small but positive $\epsilon$. Due to the symmetry of the feasible region and the posterior distribution of $\boldsymbol{\mu}|\bar{\boldsymbol{\xi}}$ around the origin, the Joint-EO solution estimator is sufficiently close to the origin, while the Separate-EO solution estimator is always on the boundary. This yields the maximum distance of the Separate-EO solution from the Joint-EO solution.

For a numerical illustration, assume that $n=2$, $\bar{\xi}_1=1.001$, $\bar{\xi}_2=-1$, $\lambda_1=-1,\lambda_2=1$, $\sigma_1=\sigma_2=\delta_1=\delta_2=1$. We compute $\eta_1=0.0005$, $\eta_2=0$ and $\zeta_1^2=\zeta_2^2=\frac{1}{2}$. It follows that the Bayes solution estimator for the Joint-EO is very close to the origin and the solution estimator for the Separate-EO is $\boldsymbol{x}=(1,0)$. ■

## Appendix D: Omitted analyses from Section 5

### D.1. Piecewise linear functions

In this section, we consider piecewise linear structures with exponential-gamma and geometric-beta conjugate pairs.

#### D.1.1. Exponential Likelihood with Gamma Prior

Using Corollary 3, one can prove the following.

PROPOSITION 12. *Consider the stochastic problem* (26). *Assume that the likelihood distribution is exponential with $\xi\sim Exp(\lambda)$ and the prior distribution is gamma with $\lambda\sim Gamma(\alpha,\beta)$. Assume further that the shape and rate hyperparameters $\alpha$ and $\beta$ are known, and a realization $\bar{\xi}$ is observed. Then, we have*

*(i)* $\hat{x}^S(\bar{\xi})=\frac{\beta+\bar{\xi}}{\alpha+1}\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)$.

*(ii)* $\hat{x}^{J,L}(\bar{\xi})=(\beta+\bar{\xi})\left[\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)^{1/(\alpha+1)}-1\right]$.

*(iii)* $\hat{x}^{J,Q}(\bar{\xi})=\frac{\beta+\bar{\xi}}{\alpha}\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)$.

*Proof:* (i) Due to Corollary 3(i), we have

$$\hat{x}^S(\bar{\xi})=G^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\hat{\lambda}^B(\bar{\xi})\right)=-\frac{1}{\hat{\lambda}^B(\bar{\xi})}\ln\left(1-\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)=\frac{1}{\hat{\lambda}^B(\bar{\xi})}\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right),$$

where $G(\xi)$ is the cdf of an exponential random variable with parameter $\hat{\lambda}^B(\bar{\xi})=\frac{\alpha+1}{\beta+\bar{\xi}}$. Hence, we obtain $\hat{x}^S(\bar{\xi})$ as stated.

(ii) Due to Corollary 3(ii), we have

$$\hat{x}^{J,L}(\bar{\xi}) = H^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\bar{\xi}\right) = \beta'\left[\left(1-\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)^{-1/\alpha'}-1\right] = \beta'\left[\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)^{1/\alpha'}-1\right],$$

where $H(\xi)$ is the cdf of a Lomax random variable with the scale and shape parameters $\beta' := \beta + \bar{\xi}$ and $\alpha' = \alpha + 1$. Hence, the result follows.

(iii) We write that

$$\begin{aligned}
\hat{x}^{J,Q}(\bar{\xi}) &= \int_0^\infty \frac{1}{\lambda}\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\Pi(\lambda|\bar{\xi})d\lambda \\
&= \ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\int_0^\infty \frac{1}{\lambda}\frac{(\beta+\bar{\xi})^{\alpha+1}}{\Gamma(\alpha+1)}\lambda^\alpha e^{-(\beta+\bar{\xi})\lambda}d\lambda \\
&= \ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)(\beta+\bar{\xi})\frac{\Gamma(\alpha)}{\Gamma(\alpha+1)}\int_0^\infty \frac{(\beta+\bar{\xi})^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-(\beta+\bar{\xi})\lambda}d\lambda \\
&= \ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\frac{(\beta+\bar{\xi})}{\alpha},
\end{aligned}$$

where the first equality follows from Corollaries 1 and 3(iii), the second equality holds since the posterior distribution $\Pi(\lambda|\bar{\xi})$ is gamma with the shape and rate parameters $\alpha+1$ and $\beta+\bar{\xi}$ respectively, the third equality is obtained by factoring suitable terms out of the integral and the last equality follows from the facts that $\frac{\Gamma(\alpha)}{\Gamma(\alpha+1)} = \frac{1}{\alpha}$ and that the integral is equal to 1 as it represents a gamma distribution. □

We note that the complexity of obtaining the Separate-EO and Joint-EO estimators is the same as they all admit closed form solutions. Next, we discuss the risk difference between $\hat{x}^S(\bar{\xi})$ and $\hat{x}^{J,L}(\bar{\xi})$ as well as $\hat{x}^{J,Q}(\bar{\xi})$.

PROPOSITION 13. *Let $\alpha > 1$. Under the assumptions of Proposition 12, we have*

$$\mathcal{R}^L\left(x^*(\mu), \hat{x}^S(\bar{\xi})\right) - \mathcal{R}^L\left(x^*(\mu), \hat{x}^{J,L}(\bar{\xi})\right) \le \frac{\tilde{b}\beta\alpha}{\alpha-1}\left\{\left[\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)^{1/(\alpha+1)}-1\right] - \frac{1}{\alpha+1}\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right\}.$$

*Proof:* Let us first compute an upper bound on the difference in the loss values of the estimators. We define $F(x) := \mathbb{E}_{\xi|\mu}[f(\xi,x)]$ as the objective function of (26) and write

$$\begin{aligned}
&\mathcal{L}^L(x^*(\theta), \hat{x}^S(\bar{\xi})) - \mathcal{L}^L(x^*(\theta), \hat{x}^{J,L}(\bar{\xi})) = F(\hat{x}^{J,L}(\bar{\xi})) - F(\hat{x}^S(\bar{\xi})) \\
\le& F'(\hat{x}^S(\bar{\xi}))(\hat{x}^{J,L}(\bar{\xi}) - \hat{x}^S(\bar{\xi})) \\
\le& \tilde{b}(\beta+\bar{\xi})\left\{\left[\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)^{1/(\alpha+1)}-1\right] - \frac{1}{\alpha+1}\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right\},
\end{aligned}$$

where the first equality follows from the definition of the linear loss (2), the first inequality is obtained from the first order Taylor expansion of the concave function $F(x)$ at point $\hat{x}^{J,L}(\bar{\xi})$ about $\hat{x}^S(\bar{\xi})$, the second inequality follows from (i) $\hat{x}^{J,L}(\bar{\xi}) - \hat{x}^S(\bar{\xi}) = (\beta+\bar{\xi})\left[\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)^{1/(\alpha+1)}-1\right] - \frac{\beta+\bar{\xi}}{\alpha+1}\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)$ due to Proposition 12, (ii) $\hat{x}^{J,L}(\bar{\xi}) > \hat{x}^S(\bar{\xi})$ since $\kappa^{1/a} - 1 > \frac{\ln(\kappa)}{a}$ for all $\kappa, a > 1$, and (iii) $F'(x) \le \tilde{b}$ for all $x \in \mathbb{R}$ which is deduced from (25). To obtain the risk difference, we take the expectations $\mathbb{E}_\lambda\mathbb{E}_{\bar{\xi}|\lambda}$ from the last term, which yields the desired result using $\mathbb{E}_\lambda\mathbb{E}_{\bar{\xi}|\lambda}[\bar{\xi}] = \frac{\beta}{\alpha-1}$. □

For the quadratic loss, it is possible to compute the risk difference exactly.

PROPOSITION 14. *Let $\alpha > 2$. Under the assumptions of Proposition 12, we have*

$$\mathcal{R}^Q\left(x^*(\mu), \hat{x}^S(\bar{\xi})\right) - \mathcal{R}^Q\left(x^*(\mu), \hat{x}^{J,Q}(\bar{\xi})\right) = \frac{\beta^2}{\alpha(\alpha+1)^2(\alpha-2)}\left[\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right]^2.$$

*Proof:* We write that

$$\begin{aligned}
\mathcal{R}^Q\left(x^*(\mu), \hat{x}^S\right) &= \left[\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right]^2 \mathbb{E}_\lambda \mathbb{E}_{\bar{\xi}|\lambda}\left[\left(\frac{1}{\lambda} - \frac{\beta+\bar{\xi}}{\alpha+1}\right)^2\right] \\
&= \left[\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right]^2 \mathbb{E}_\lambda \mathbb{E}_{\bar{\xi}|\lambda}\left[\frac{1}{\lambda^2} - \frac{2}{\lambda}\frac{\beta+\bar{\xi}}{\alpha+1} + \frac{(\beta+\bar{\xi})^2}{(\alpha+1)^2}\right] \\
&= \left[\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right]^2 \mathbb{E}_\lambda\left[\frac{1}{\lambda^2} - \frac{2}{\lambda}\frac{\beta+\frac{1}{\lambda}}{\alpha+1} + \frac{\beta^2+2\frac{\beta}{\lambda}+\frac{2}{\lambda^2}}{(\alpha+1)^2}\right] \\
&= \left[\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right]^2 \left\{\mathbb{E}_\lambda\left[\frac{1}{\lambda^2}\right] - \frac{2}{\alpha+1}\mathbb{E}_\lambda\left[\frac{\beta}{\lambda}+\frac{1}{\lambda^2}\right] + \frac{1}{(\alpha+1)^2}\left(\beta^2 + 2\mathbb{E}_\lambda\left[\frac{\beta}{\lambda}+\frac{1}{\lambda^2}\right]\right)\right\},
\end{aligned}$$

where the first equality follows from Proposition 12 and the definition (6) of the risk under quadratic loss, and the second equality holds because $\mathbb{E}_{\xi|\lambda}[\xi] = \frac{1}{\lambda}$ and $\mathbb{E}_{\xi|\lambda}[\xi^2] = \frac{2}{\lambda^2}$. Using similar arguments, we can compute

$$\mathcal{R}^Q\left(x^*(\mu), \hat{x}^{J,Q}(\bar{\xi})\right) = \left[\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right]^2 \left\{\mathbb{E}_\lambda\left[\frac{1}{\lambda^2}\right] - \frac{2}{\alpha}\mathbb{E}_\lambda\left[\frac{\beta}{\lambda}+\frac{1}{\lambda^2}\right] + \frac{1}{\alpha^2}\left(\beta^2 + 2\mathbb{E}_\lambda\left[\frac{\beta}{\lambda}+\frac{1}{\lambda^2}\right]\right)\right\}.$$

Note that $\lambda \sim \text{Gamma}(\alpha, \beta)$ which implies that $\frac{1}{\lambda} \sim \text{Inverse-Gamma}(\alpha, \beta)$. Therefore, we obtain that $\mathbb{E}_\lambda[\frac{1}{\lambda}] = \frac{\beta}{\alpha-1}$ and $\mathbb{E}_\lambda[\frac{1}{\lambda^2}] = \frac{\beta^2}{(\alpha-1)(\alpha-2)}$, which yields $\mathbb{E}_\lambda[\frac{\beta}{\lambda} + \frac{1}{\lambda^2}] = \frac{\beta^2}{\alpha-2}$. Combining the above results, we obtain

$$\begin{aligned}
&\mathcal{R}^Q\left(x^*(\mu), \hat{x}^S(\bar{\xi})\right) - \mathcal{R}^Q\left(x^*(\mu), \hat{x}^{J,Q}(\bar{\xi})\right) \\
&= \left[\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right]^2 \left\{2\left(\frac{1}{\alpha} - \frac{1}{\alpha+1}\right)\frac{\beta^2}{\alpha-2} + \left(\frac{1}{(\alpha+1)^2} - \frac{1}{\alpha^2}\right)\left(\beta^2 + \frac{2\beta^2}{\alpha-2}\right)\right\} \\
&= \left[\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right]^2 \left\{\frac{2}{\alpha(\alpha+1)}\frac{\beta^2}{\alpha-2} + \frac{-2\alpha-1}{\alpha^2(\alpha+1)^2}\frac{\beta^2\alpha}{\alpha-2}\right\} \\
&= \left[\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)\right]^2 \frac{\beta^2}{\alpha(\alpha+1)(\alpha-2)}\left(2 - \frac{2\alpha+1}{\alpha+1}\right),
\end{aligned}$$

from which the result follows. □

**D.1.2. Geometric Likelihood with Beta Prior** Corollary 3 implies the following.

PROPOSITION 15. *Consider the stochastic problem* (26). *Assume that the likelihood distribution is geometric with $\xi \sim Geo(p)$ and the prior distribution is beta with $p \sim Beta(\alpha, \beta)$. Assume further that the shape parameters $\alpha$ and $\beta$ are known, and a realization $\bar{\xi}$ is observed from the likelihood. Then, we have*

*(i)* $\hat{x}^S(\bar{\xi}) \approx \frac{\ln\left(\frac{\bar{a}-\tilde{a}-\tilde{b}}{\bar{a}-\tilde{a}}\right)}{\ln[(\beta+\bar{\xi}-1)/(\alpha+\beta+\bar{\xi})]}$.

*(ii)* $\hat{x}^{J,L}(\bar{\xi}) \approx \max\left\{x : \sum_{\xi=0}^{x} \frac{\alpha+1}{\alpha+\beta+\xi+2}\ \frac{\beta+\bar{\xi}+\xi}{\alpha+\beta+\xi+1}\ \frac{\beta+\bar{\xi}+\xi-1}{\alpha+\beta+\bar{\xi}+\xi} \le \frac{\tilde{b}}{\bar{a}-\tilde{a}}\right\}$.

*(iii)* $\hat{x}^{J,Q}(\bar{\xi}) \approx \frac{\ln\left(\frac{\bar{a}-\tilde{a}-\tilde{b}}{\bar{a}-\tilde{a}}\right)}{B(\alpha+1,\beta+\bar{\xi}-1)} \int_0^1 \frac{p^\alpha(1-p)^{\beta+\bar{\xi}-2}}{\ln(1-p)} dp$.

*Proof:* (i) Due to Corollary 3(i), we have

$$\hat{x}^S(\bar{\xi}) = G^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\hat{p}^B(\bar{\xi})\right) \approx \frac{\ln\left(\frac{\bar{a}-\tilde{a}-\tilde{b}}{\bar{a}-\tilde{a}}\right)}{\ln(1-\hat{p}^B(\bar{\xi}))},$$

where $G(\xi)$ is the cdf of a geometric random variable with parameter $\hat{p}^B(\bar{\xi}) = \frac{\alpha+1}{\alpha+\beta+\bar{\xi}}$. Hence, we obtain $\hat{x}^S(\bar{\xi})$ as stated.

(ii) Due to Corollary 3(ii), we have

$$\hat{x}^{J,L}(\bar{\xi}) = H^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\bar{\xi}\right),$$

where $H$ is the cdf of posterior predictive distribution. The results follows due to the relationship between the cdf $H$ and its pmf $h$.

(iii) We write that

$$\hat{x}^{J,Q}(\bar{\xi}) \approx \int_0^1 \frac{\ln\left(\frac{\bar{a}-\tilde{a}-\tilde{b}}{\bar{a}-\tilde{a}}\right)}{\ln(1-p)}\Pi(p|\bar{\xi})dp = \ln\left(\frac{\bar{a}-\tilde{a}-\tilde{b}}{\bar{a}-\tilde{a}}\right)\int_0^1 \frac{1}{\ln(1-p)}\frac{p^{\alpha}(1-p)^{\beta+\bar{\xi}-2}}{\mathrm{B}(\alpha+1,\beta+\bar{\xi}-1)}dp,$$

where the first relation follows from Corollaries 1 and 3(iii) and the second relation holds since the posterior distribution $\Pi(p|\bar{\xi})$ is beta with parameters $\alpha+1$ and $\beta+\bar{\xi}-1$. □

We note that $\hat{x}^S(\bar{\xi})$ can be approximated by a closed form expression while the approximations of $\hat{x}^{J,L}(\bar{\xi})$ and $\hat{x}^{J,Q}(\bar{\xi})$ require an algorithm and numerical integration, in general. This is an instance where computing the Joint-EO solutions is harder than computing the Separate-EO solutions.

### D.2. Sum of piecewise linear functions

In this section, we consider the sum of piecewise linear functions with exponential-gamma conjugate pairs.

Proposition 16. *Consider the one-dimensional stochastic median problem. Assume that, for each $i \in [n]$, the likelihood distribution is exponential with $\xi_i \sim Exp(\lambda_i)$ and the prior distribution is gamma with $\lambda_i \sim Gamma(\alpha_i, \beta_i)$. Assume further that the parameters $\alpha_i$ and $\beta_i$ are known with $\alpha_i > 2$, and that a realization of locations $\bar{\xi}_i$ is observed for $i \in [n]$. Then, we have*

$$\mathcal{R}^L\left(x^*(\boldsymbol{\lambda}), \hat{x}^S(\bar{\xi})\right) - \mathcal{R}^L\left(x^*(\boldsymbol{\lambda}), \hat{x}^{J,L}(\bar{\xi})\right) \le n\Bigg(\max_i\left\{\frac{\alpha_i\beta_i(\sqrt[\alpha_i+1]{2}-1)}{\alpha_i-1}\right\}$$
$$+\sqrt{\frac{n-1}{n}\sum_{i=1}^n \frac{\beta_i^2\alpha_i(\sqrt[\alpha_i+1]{2}-1)}{(\alpha_i-1)^2(\alpha_i-2)}} - \min_i\left\{\frac{\alpha_i\beta_i\ln 2}{\alpha_i-1}\right\} + \sqrt{\frac{n-1}{n}\sum_{i=1}^n \frac{\beta_i^2\alpha_i(\ln 2)^2}{(\alpha_i-1)^2(\alpha_i-2)(\alpha_i+1)^2}}\Bigg).$$

*Proof:* Proposition 12 implies that $\hat{x}_i^S(\bar{\xi}_i) = \frac{\beta_i+\bar{\xi}_i}{\alpha_i+1}\ln 2$ and $\hat{x}_i^{J,L}(\bar{\xi}_i) = (\beta_i+\bar{\xi}_i)(\sqrt[\alpha_i+1]{2}-1)$ as $\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}} = 2$ for $i \in [n]$. Since $\frac{\ln 2}{\alpha_i+1} < \sqrt[\alpha_i+1]{2}-1$ for $\alpha_i > 0$, we have $\hat{x}_i^S(\bar{\xi}_i) < \hat{x}_i^{J,L}(\bar{\xi}_i)$ for each $i \in [n]$. Applying Proposition 10, we obtain

$$\mathcal{R}^L\left(x^*(\boldsymbol{\lambda}), \hat{x}^S(\bar{\xi})\right) - \mathcal{R}^L\left(x^*(\boldsymbol{\lambda}), \hat{x}^{J,L}(\bar{\xi})\right) \le n\mathbb{E}_{\bar{\boldsymbol{\xi}}}[\max_i\{\hat{x}_i^{J,L}(\bar{\xi}_i)\} - \min_i\{\hat{x}_i^S(\bar{\xi}_i)\}]. \tag{39}$$

Now, we compute an upper bound on the right-hand-side using Aven (1985). Since $\bar{\xi}_i \sim \mathrm{Lomax}(\beta_i, \alpha_i)$ for $i \in [n]$, we have

$$\mathbb{E}[\xi_i] = \frac{\beta_i}{\alpha_i-1} \text{ and } \mathrm{Var}(\xi_i) = \frac{\beta_i^2\alpha_i}{(\alpha_i-1)^2(\alpha_i-2)}.$$

Since both $\hat{x}_i^S(\bar{\xi}_i)$ and $\hat{x}_i^{J,L}(\bar{\xi}_i)$ are affine transformations of $\xi_i$, we can easily obtain their mean and variance as well. Finally, plugging in the resulting bounds for $\mathbb{E}[\min_i\{\hat{x}_i^S(\bar{\xi}_i)\}]$ and $\mathbb{E}[\max_i\{\hat{x}_i^{J,L}(\bar{\xi}_i)\}]$ into (39) gives the desired result. □

## Acknowledgments

This work was supported in parts by ONR grant 000141812129 and NSF grant CMMI-1560828.

## Author Biographies

**Danial Davarnia** is an Assistant Professor in the Department of Industrial and Manufacturing Systems Engineering at Iowa State University. He received his Ph.D. degree in Industrial and Systems Engineering from University of Florida, and was a postdoctoral research fellow in the Tepper School of Business at Carnegie Mellon University after that. His research interests include mixed-integer nonlinear programming and stochastic programming, with applications in transportation and network design.

**Burak Kocuk** is an Assistant Professor in the Industrial Engineering Program at Sabancı University, Istanbul. He received his Ph.D. degree in Operation Research from Georgia Institute of Technology and he was a postdoctoral research fellow at Carnegie Mellon University after that. His research focuses on developing solution methods for mixed-integer nonlinear programming problems with applications in engineering optimization.

**Gérard Cornuéjols** is a University Professor at Carnegie Mellon University. He holds the IBM chair in Operations Research. He is a member of the National Academy of Engineering; he was awarded the Lanchester prize, the Fulkerson prize, the von Neumann theory prize and the Dantzig prize.

**Story of the paper**.

Practitioners in the finance industry have observed a fascinating interaction between data and optimization. In the classical Markowitz model for portfolio construction, a major issue is the estimation of the expected returns of the individual assets. The theory of efficient markets tells us that the latest data contain the best available information. Surprisingly, in 1986 Jorion recommended not to use the data directly, but instead to "shrink" the data on individual assets towards a "grand average" involving all the other assets as well. This seems counterintuitive but it works well in practice. This "paradox" can be traced back to the stunning work of Stein (1956) in statistics. Our work originated in an attempt to understand the computational consequences of this paradox in the more general context of optimization.

## References

Aven T (1985) Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables. *Journal of Applied Probability* 22(3):723–728.

Basu A, Nguyen T, Sun A (2019) Admissibility of solution estimators for stochastic optimization. *arXiv preprint arXiv:1901.06976* .

Ben-Tal A, Ghaoui L, Nemirovski A (2009) *Robust Optimization* (Princeton Series in Applied Mathematics).

Bertsimas D, Brown DB, Caramanis C (2011) Theory and applications of robust optimization. *SIAM Review* 53:464–501.

Bertsimas D, Gupta V, Paschalidis I (2012) Inverse optimization: A new perspective on the Black-Litterman model. *Operations Research* 60(6):1389–1403.

Birge J (2016) Uses of sub-sample estimates in stochastic optimization models. *Operations Research* .

Birge J, Louveaux F (2011) *Introduction to stochastic programming* (Springer).

Black F, Litterman R (1991) Asset allocation: Combining investor views with market equilibrium. *Journal of Fixed Income* 1(2):7–18.

Black F, Litterman R (1992) Global portfolio optimization. *Financial Analysts Journal* 48(5):28–43.

Boyd S, Kim SJ, Vandenberghe L, Hassibi A (2007) A tutorial on geometric programming. *Optimization and Engineering* 8:67–127.

Chu LY, Shanthikumar JG, Shen ZJM (2008) Solving operational statistics via a Bayesian analysis. *Operations Research Letters* 36:110–116.

Davarnia D, Cornuéjols G (2017) From estimation to optimization via shrinkage. *Operations Research Letters* 45:642–646.

Diaconis P, Ylvisaker D (1979) Conjugate priors for exponential families. *Annals of Statistics* 7:269–281.

Donti P, Amos B, Kolter Z (2017) Task-based end-to-end model learning. `https://arxiv.org/abs/1703.04529v1` .

Ferguson T (1967) *Mathematical Statistics-A Decision Theoretic Approach* (New York and London: Academic Press).

Goodfellow I, Bengio Y, Courville A (2016) *Deep Learning* (Cambridge, MA: The MIT Press).

Jiang H, Shanbhag UV (2016) On the solution of stochastic optimization and variational problems in imperfect information regimes. *SIAM Journal on Optimization* 24:2394–2429.

Jorion P (1986) Bayes-Stein estimation for portfolio analysis. *The Journal of Financial and Quantitative Analysis* 21:279–292.

Kadane JB (2011) *Principles of Uncertainty* (Chapman and Hall, CRC Press).

Kleywegt AJ, Shapiro A (2004) Stochastic optimization. Technical report, Georgia Institute of Technology.

Kocuk B, Cornuéjols G (2018) Incorporating Black-Litterman views in portfolio construction when stock returns are a mixture of normals. *Omega* URL `http://dx.doi.org/10.1016/j.omega.2018.11.017`.

Kohli N, Harring JR (2013) Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research* 48:370–397.

Lehmann EL, Casella G (1998) *Theory of Point Estimation* (Springer-Verlag).

Levine S, Finn C, Darrell R, Abbeel P (2016) End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research* 17:1–40.

Lim AEB, Shanthikumar JG, Shen ZJM (2006) Model uncertainty, robust optimization, and learning. Johnson MP, Norman B, Secomandi N, eds., *Tutorials in Operations Research: Models, Methods, and Applications for Innovative Decision Making*, 66–94 (INFORMS).

Liyanage LH, Shanthikumar JG (2005) A practical inventory control policy using operational statistics. *Operations Research Letters* 33:341–348.

Markowitz HM (1959) *Portfolio Selection: Efficient Diversification of Investments* (John Wiley).

Shapiro A (2003) Monte Carlo sampling methods. Ruszczyński A, Shapiro A, eds., *Stochastic Programming, Handbooks in Operations Research and Management Science*, volume 10, 352–425 (Elsevier).

Stein C (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 197–206 (University of California Press).

Thomas RW, Friend DH, Dasilva LA, Mackenzie AB (2006) Cognitive networks: adaptation and learning to achieve end-to-end performance objectives. *IEEE Communications Magazine* 44:51–57.

Vielma JP, Ahmed S, Nemhauser G (2010) Mixed-integer models for nonseparable piecewise linear optimization: unifying framework and extensions. *Operations Research* 58:303–315.

Wang M, Fang EX, Liu H (2017) Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming* 161:419–449.