BAYESIAN SOLUTION ESTIMATORS IN STOCHASTIC OPTIMIZATION*

DANIAL DAVARNIA†, BURAK KOCUK‡, AND GÉRARD CORNUÉJOLS§

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 natural 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 study 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. In particular, (i) we identify conditions under which the solution estimators of both schemes are equal, (ii) for general problems, we show that the risk difference between the two schemes can be arbitrarily large, (iii) for stochastic piecewise linear programs, we derive explicit bounds on risk differences, and (iv) for stochastic geometric programs, we discuss the difference in computational complexity of the two schemes and provide computational experiments.

Key words. Stochastic optimization, Bayesian inference, Estimation, Piecewise linear function, Geometric program

AMS subject classifications. 90C15, 62C10, 62F10

1. Introduction.

Consider the following stochastic program

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

In (1), the vector $\boldsymbol{x} \in \mathbb{R}^n$ represents random variables with joint probability density function $g(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes a vector of parameters. Vector $\boldsymbol{y} \in \mathbb{R}^m$ represents decision variables that must belong to a closed set $\mathcal{Y} \subseteq \mathbb{R}^m$. Function $f(\boldsymbol{x},\boldsymbol{y}) : \mathbb{R}^{n+m} \to \mathbb{R}$ contains both random variables and decision variables, and its expectation with respect to the distribution $g(\boldsymbol{x}|\boldsymbol{\theta})$ forms the objective function of the stochastic program.

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

---

*Submitted to the editors DATE.

Funding: This work was supported in parts by ONR grant 00014-15-12082 and NSF grant CMMI-1560828.

†Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA (ddavarni@andrew.cmu.edu).

‡Industrial Engineering Program, Sabancı University, Istanbul, Turkey (burakkocuk@sabanciuniv.edu).

§Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA (gc0v@andrew.cmu.edu).

Optimization models where some elements (such as coefficients) arise from a random process are often formulated as stochastic programs. We refer the reader to [6] for an introduction to stochastic programming, and to [31, 32] for applications in energy and finance. Various solution techniques have been proposed in the literature for different classes of stochastic programs. Examples of such techniques are Monte Carlo, bootstrapping and scenario reduction techniques such as sample-average approximation (SAA); see [26]. Convergence results for the SAA method are discussed in [27, 28], while bounds on the number of samples required to achieve converging performance are derived in [8, 18, 29] . We refer the reader to [21] for convergence results for the newsvendor problem and to [12] for the minimum spanning tree with applications in network design.

Point estimation has been vastly studied in statistical inference and decision theory. We refer the interested reader to [20] and many other references therein. Estimation plays a central role in constructing appropriate optimization models that involve uncertainty. For instance, the theory of minimaxity in estimation is closely related to robust optimization. We refer the reader to [2, 3, 4] for introduction to the theory and applications of robust optimization, and to [23] and reference therein for a detailed account of the connection between robust optimization and minimaxity.

Several contributions evaluate the impact of integrating estimation and optimization from various perspectives. In [10], the authors study a class of quadratic stochastic programs and investigate the quality of the popular maximum likelihood estimator. A recent study in the area of machine learning by [14] focuses on training predictive models whose outcome is used in stochastic programs. Other related lines of work in machine learning explore the end-to-end training systems where the end goal is directly predicted from the raw input; see [22, 30] for instance. [9, 24] study a single-period univariate newsvendor problem and introduce the notion of *operational statistics* as a scheme that uses the optimization structure to design desirable estimators. Another area where the combination of estimation and optimization can lead to a better outcome is concerned with iterative estimation and optimization models; see [17, 16] for a detained account.

In this paper, we focus on stochastic optimization and its connection to statistical estimation theory. There are two main factors to fix before initiating this investigation: (i) the criteria to measure the quality of solution estimators, (ii) the schemes to compute solution estimators. For (i), we consider a standard statistical criterion (called *risk*) defined based on two natural measures in optimization: one that considers the average difference between the objective value of the solution estimator and the true optimal value of the stochastic problem, and the other that considers the average Euclidean distance between the solution estimator and the true optimal solution. For (ii), we consider two natural schemes to compute solution estimators. In the first scheme, the estimation process is performed separately from the optimization, and its output estimator is used as a known input for the optimization problem. In the second scheme, the estimation process is incorporated within the optimization step to obtain a solution. The key question in this context is how different the risk value and the computational complexity of the estimators obtained from the separate and joint schemes are, based on the two criteria defined above?

We address this question from several angles under a Bayesian framework, where a prior distribution is assumed for the unknown parameters. In Section 2, we formally define the risk criteria and the separate and joint estimation-optimization schemes. In Section 3, we identify conditions under which the two schemes yield the same solutions, and provide examples that show an arbitrary large risk difference between the two schemes when these conditions do not hold. In Section 4, we study a class of stochastic piecewise linear programs and evaluate the risk difference between the separate and joint schemes both theoretically and computationally. This class generalizes the newsvendor and the median problems, and models several applications in education and psychology. In Section 5, we conduct a risk assessment analysis for the stochastic geometric programs. This structure has applications in optimal control, network design and chemical equilibrium. We observe that the separate and joint schemes lead to problems with different computational complexity for certain distributions. We provide computational results to support this assertion. Section 6 contains concluding remarks.

2. Loss Function and Risk. Consider the stochastic program (1). Assume that variables $\boldsymbol{x}$ have a joint distribution $g(\boldsymbol{x}|\boldsymbol{\theta})$, where parameters $\boldsymbol{\theta}$ are unknown. Throughout this paper, we assume that a single observation sampled from this distribution is available, i.e., $T = 1$. This will simplify our notation while preserving the richness of the setting. We denote the vector of random variables corresponding to this single observation by $\bar{\boldsymbol{x}}$.

Since the true optimal solution $\boldsymbol{y}^*(\boldsymbol{\theta})$ is unknown (as $\boldsymbol{\theta}$ is unknown), we estimate it with a *solution estimator* $\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}}) \in \mathcal{Y}$ as a vector function of the observation $\bar{\boldsymbol{x}}$. There are many choices for a solution estimator, hence we need a suitable criterion to evaluate its performance. We use *risk* as a popular criterion in evaluating the performance of estimators in statistical inference. To this end, we define a *loss* function that measures the difference or distance between solutions we intend to compare. In the next two subsections, we define two natural loss functions for optimization problems, 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 quality of solution estimators is measured by a *loss function*. In this section, the loss function is defined as the difference between the objective values of the solution estimator $\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})$ and the true optimal solution $\boldsymbol{y}^*(\boldsymbol{\theta})$.

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

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

Throughout this paper, 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 $\mathcal{L}^L\left(\boldsymbol{y}^*(\boldsymbol{\theta}), \hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})\right)$ is a function of both the unknown parameters $\boldsymbol{\theta}$ and the observation $\bar{\boldsymbol{x}}$, it is a random quantity itself due to the randomness in these two elements. To obtain a measure of overall performance for such a random quantity, a *risk* is defined as

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

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 distribution $g(\bar{\boldsymbol{x}}|\boldsymbol{\theta})$ of observation $\bar{\boldsymbol{x}}$ given $\boldsymbol{\theta}$.

With the above definition of risk, a *best* solution estimator is defined as one that achieves minimum risk. We refer to such estimators as *Bayes solution estimators*.

PROPOSITION 2.1. *Consider stochastic program* (1). *Assume that* $\boldsymbol{x} \sim g(\boldsymbol{x}|\boldsymbol{\theta})$ *and* $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$. *Then, any solution estimator* $\hat{\boldsymbol{y}}^{J,L}(\bar{\boldsymbol{x}})$ *that solves*

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

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

The proofs of the propositions in Sections 2 and 3 are given in Appendix A. Proposition 2.1 gives a joint estimation and optimization (Joint-EO) scheme that solves the estimation problem simultaneously with the optimization problem to obtain a solution estimator with minimum risk for the loss function (2). We denote this Bayes solution estimator by $\hat{\boldsymbol{y}}^{J,L}(\bar{\boldsymbol{x}})$, where the superscript represents *Joint* method with respect to the *Linear* loss (2).

While this joint scheme yields the best solution, traditionally, the estimation and optimization problems are solved separately. The estimation problem in this separate scheme does not use any information about the structure of the optimization problem, thus it requires a different loss function to evaluate the risk. The most common loss function used in statistical point estimation is the squared error loss $\mathcal{L}^Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\bar{\boldsymbol{x}})) = ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}(\bar{\boldsymbol{x}})||^2$ which measures the Euclidean distance of the estimator $\hat{\boldsymbol{\theta}}(\bar{\boldsymbol{x}})$ from the true unknown parameter $\boldsymbol{\theta}$; see [15]. It can be shown that the Bayes estimator of $\boldsymbol{\theta}$ under the squared error loss $\mathcal{L}^Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\bar{\boldsymbol{x}}))$ is $\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{x}}) = \mathbb{E}_{\theta|\bar{x}}[\boldsymbol{\theta}]$, where the expectation is taken with respect to the posterior distribution $\Pi(\boldsymbol{\theta}|\bar{\boldsymbol{x}})$ of the parameters $\boldsymbol{\theta}$ given the observation $\bar{\boldsymbol{x}}$. A proof for a more general case is given in Corollary 2.6.

This result sets up the basis for a separate estimation and optimization (Separate-EO) scheme that solves the estimation problem independently of the optimization problem. First, a Bayes estimator $\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{x}})$ for $\boldsymbol{\theta}$ is computed with respect to the squared error loss, then this estimator is used in place of $\boldsymbol{\theta}$ in (1) to obtain an optimal solution $\hat{\boldsymbol{y}}^S(\bar{\boldsymbol{x}})$ that serves as a solution estimator for the true optimal solution $\boldsymbol{y}^*(\boldsymbol{\theta})$. The superscript in $\hat{\boldsymbol{y}}^S(\bar{\boldsymbol{x}})$ represents the *Separate* scheme. It follows that the Separate-EO solution estimator $\hat{\boldsymbol{y}}^S(\bar{\boldsymbol{x}})$ may be suboptimal in risk to the Joint-EO solution estimator $\hat{\boldsymbol{y}}^{J,L}(\bar{\boldsymbol{x}})$, as the former is a feasible solution of (4) while the latter is an optimal solution.

Remark 2.2. When the distribution of $\boldsymbol{x}$ belongs to an exponential family, it admits a *conjugate* prior, leading to a posterior distribution of the same family as the prior. Furthermore, for such distributions, the posterior mean (Bayes estimator of the parameter) is representable as a convex combination between the prior mean and the observation (MLE for multiple observations); see [13]. Such estimators are sometimes referred to as *shrinkage* estimators as they shrink the MLE towards another vector. For instance, assume that $x_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ for $i \in [n]$ and $\mu_i \sim \mathcal{N}(\lambda_i, \delta_i^2)$ and all variables are independent. Then, the Bayes estimator $\hat{\boldsymbol{\mu}}^B(\bar{\boldsymbol{x}})$ for $\boldsymbol{\mu}$ is such that $\hat{\mu}_i^B = \rho_i \lambda_i + (1 - \rho_i)\bar{x}_i$ where $\rho_i = \frac{\sigma_i^2}{\sigma_i^2 + \delta_i^2}$ for $i \in [n]$.

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

Definition 2.3. Assume that $\boldsymbol{x} \sim g(\boldsymbol{x}|\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$, and that a single observation $\bar{\boldsymbol{x}}$ is sampled from the distribution of $\boldsymbol{x}$. Then, the *posterior predictive* distribution of $\boldsymbol{x}$ given $\bar{\boldsymbol{x}}$ is defined as $h(\boldsymbol{x}|\bar{\boldsymbol{x}}) = \int_{\theta} g(\boldsymbol{x}|\boldsymbol{\theta})\Pi(\boldsymbol{\theta}|\bar{\boldsymbol{x}})d\boldsymbol{\theta}$, which is obtained by marginalizing the distribution of $\boldsymbol{x}$ given $\boldsymbol{\theta}$ over the posterior distribution of $\boldsymbol{\theta}$ given observation $\bar{\boldsymbol{x}}$. The definition for discrete distributions is similar.

Using Definition 2.3, we obtain an equivalent expression for the Joint-EO problem (4):

PROPOSITION 2.4. *Assume that $\boldsymbol{x} \sim g(\boldsymbol{x}|\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$. Then, any solution estimator $\hat{\boldsymbol{y}}^{J,L}(\bar{\boldsymbol{x}})$ that solves*

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

is a Bayes solution estimator under the loss function (2), where the expectation is taken with respect to the posterior predictive distribution $h(\boldsymbol{x}|\bar{\boldsymbol{x}})$ of the random variables $\boldsymbol{x}$ given the observation $\bar{\boldsymbol{x}}$.

Proposition 2.4 allows for a direct comparison between the problems to solve for 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 $g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{x}}))$ of $\boldsymbol{x}$ given $\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{x}})$, which is the Bayes estimator for the unknown parameters $\boldsymbol{\theta}$ given the observation $\bar{\boldsymbol{x}}$. For the Joint-EO under the loss (2), we solve (5) where the expectation is taken with respect to the posterior predictive distribution $h(\boldsymbol{x}|\bar{\boldsymbol{x}})$ of $\boldsymbol{x}$ given observation $\bar{\boldsymbol{x}}$. Therefore, the difference between these two estimation schemes stems from the difference between the distributions with respect to which the expectation of the objective function $f(\boldsymbol{x},\boldsymbol{y})$ is computed.

2.2. Loss as the distance between solutions. In the previous section, we defined the loss function as the difference between objective values of the solution estimator and the true optimal solution. Another natural loss function for optimization problems is the (Euclidean) distance between the solution estimator and the optimal solution. Since an optimization problem can have multiple optimal solutions, we first define the distance of a point from a set as follows

$$\mathcal{D}(W,v) = \min_{w\in W} ||w-v||, \tag{6}$$

where $W \subseteq \mathbb{R}^n$ and $v \in \mathbb{R}^n$. Using this distance measure, we define the loss function

$$\mathcal{L}^Q\left(\boldsymbol{y}^*(\boldsymbol{\theta}), \hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})\right) = \mathcal{D}^2(\boldsymbol{y}^*(\boldsymbol{\theta}), \hat{\boldsymbol{y}}(\bar{\boldsymbol{x}}))\,, \tag{7}$$

which measures the Euclidean distance between the solution estimator $\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})$ and the set $\boldsymbol{y}^*(\boldsymbol{\theta})$ of optimal solutions of (1) when $\boldsymbol{\theta}$ is known. The superscript represents that the loss is *quadratic* in solutions. We define risk similarly to (3) as the expectation of (7) with respect to the distribution $g(\boldsymbol{x}|\boldsymbol{\theta})$ and the prior $\pi(\boldsymbol{\theta})$:

$$\mathcal{R}^Q\left(\boldsymbol{y}^*(\boldsymbol{\theta}), \hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})\right) = \mathbb{E}_{\theta}\mathbb{E}_{\bar{x}|\theta}\left[\mathcal{L}^Q\left(\boldsymbol{y}^*(\boldsymbol{\theta}), \hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})\right)\right]. \tag{8}$$

The goal here is to investigate how this change of the loss function affects the Separate-EO and Joint-EO methods introduced in Section 2.1. 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 $\hat{\boldsymbol{y}}^S(\bar{\boldsymbol{x}})$ as given in Section 2.1. For the Joint-EO method, however, the optimization problem to solve is different and therefore a different solution estimator is obtained.

PROPOSITION 2.5. *Assume that $\boldsymbol{x} \sim g(\boldsymbol{x}|\boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$. Then, any solution estimator $\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}})$ that solves*

$$\min_{\boldsymbol{y}\in\mathcal{Y}} \; \mathbb{E}_{\theta|\bar{x}}\left[\mathcal{D}^2(\boldsymbol{y}^*(\boldsymbol{\theta}), \boldsymbol{y})\right], \tag{9}$$

is a Bayes solution estimator under the loss function (7), where the expectation is taken with respect to the posterior distribution $\Pi(\boldsymbol{\theta}|\bar{\boldsymbol{x}})$ of the parameters $\boldsymbol{\theta}$ given the observation $\bar{\boldsymbol{x}}$.

In Proposition 2.5, $\mathcal{D}^2(\boldsymbol{y}^*(\boldsymbol{\theta}), \boldsymbol{y})$ is a function of $\boldsymbol{\theta}$ which measures the distance of a point $\boldsymbol{y}$ to the set $\boldsymbol{y}^*(\boldsymbol{\theta})$ of optimal solutions of (1), and the optimization problem tries to find the Joint-EO solution estimator $\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}})$ as a minimizer of the posterior expectation of this distance. The superscript in $\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}})$ denotes the *Joint* method under the *Quadratic* loss function (7). The distance function $\mathcal{D}^2(\boldsymbol{y}^*(\boldsymbol{\theta}), \boldsymbol{y})$ can be very complicated especially when the set $\boldsymbol{y}^*(\boldsymbol{\theta})$ of optimal solutions contains several points. This already implies that solving the Joint-EO problem under the loss (7) can be much harder than the solution estimators obtained from the Separate-EO problem or the Joint-EO problem under the loss (2) as discussed in Section 2.1. Note that the multiplicity of optimal solutions does not create an issue for the loss function (2) as it only contains the optimal value (which is unique).

To streamline the discussion for the Joint-EO problem (9), we assume that the optimal solution of (1) is unique and therefore $\boldsymbol{y}^*(\boldsymbol{\theta})$ is a single point. In this case, the distance function (6) reduces to the Euclidean norm, yielding the loss function $\mathcal{L}^Q\left(\boldsymbol{y}^*(\boldsymbol{\theta}), \hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})\right) = ||\boldsymbol{y}^*(\boldsymbol{\theta}) - \hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})||^2$. Therefore, we can solve (9) in a reduced-form as follows. In the sequel, the expectation operator $\mathbb{E}[.]$ applied on a vector implies component-wise expectation, i.e., $\mathbb{E}[\boldsymbol{v}] = \boldsymbol{w}$ where $w_i = \mathbb{E}[v_i]$ for all $i \in [n]$.

COROLLARY 2.6. *Assume that (1) has a unique optimal solution $\boldsymbol{y}^*(\boldsymbol{\theta})$ for any given $\boldsymbol{\theta}$. Then, the Bayes solution estimator of (9) under the loss function (7) is obtained as the optimal solution of*

$$\min_{\boldsymbol{y}\in\mathcal{Y}} \; ||\mathbb{E}_{\theta|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\theta})] - \boldsymbol{y}||^2. \tag{10}$$

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

Proof. Since $\boldsymbol{y}^*(\boldsymbol{\theta})$ is unique, we obtain that $\mathcal{D}^2(\boldsymbol{y}^*(\boldsymbol{\theta}), \hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})) = ||\boldsymbol{y}^*(\boldsymbol{\theta}) - \hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})||^2$ by definition (6). To obtain the Bayes solution estimator, we need to find a minimizer of (9) which reduces to $\min_{\boldsymbol{y}\in\mathcal{Y}} \; \mathbb{E}_{\theta|\bar{x}}\left[||\boldsymbol{y}^*(\boldsymbol{\theta}) - \boldsymbol{y}||^2\right]$. Define $\boldsymbol{w} = \mathbb{E}_{\theta|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\theta})]$. We write that

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

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 $\mathbb{E}_{\theta|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\theta}) - \boldsymbol{w}] = \boldsymbol{0}$ by definition, and because $||\boldsymbol{w} - \boldsymbol{y}||^2$ does not depend on $\boldsymbol{\theta}$. Note in the last relation that the first term $\mathbb{E}_{\theta|\bar{x}}\left[||\boldsymbol{y}^*(\boldsymbol{\theta}) - \boldsymbol{w}||^2\right]$ does not contain $\boldsymbol{y}$. This gives the desired relation (10). For the next result, when $\mathcal{Y}$ is convex, any convex combination of $\boldsymbol{y}^*(\boldsymbol{\theta})$ belongs to $\mathcal{Y}$ as $\boldsymbol{y}^*(\boldsymbol{\theta}) \in \mathcal{Y}$. We conclude that $\mathbb{E}_{\theta|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\theta})] \in \mathcal{Y}$, and therefore the minimizer of (10) is attained at $\mathbb{E}_{\theta|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\theta})]$. ∎

In Corollary 2.6, the uniqueness of the optimal solution converts the joint estimation problem to a *projection* problem. In particular, the Bayes solution is the projection of the expected optimal solution $\mathbb{E}_{\theta|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\theta})]$ onto the feasible region $\mathcal{Y}$. In this case, the convexity of $\mathcal{Y}$ is sufficient to guarantee that the expected optimal solution belongs to the feasible region, and therefore this expected solution is the Bayes solution.

3. General Problem Structures. In Sections 2.1 and 2.2, we introduced two natural loss functions for optimization problems as well as two common estimation-schemes to obtain solution estimators. Because there is a trade-off between risk performance and calculation effort, two basic questions arise: (i) Under what conditions do the Separate-EO and Joint-EO methods lead to the same estimator? (ii) When the two estimators are different, is there a bound on their risk differences?

3.1. When the separate and joint estimators are the same. In this section, we identify sufficient conditions under which the solution estimators obtained from the Separate-EO and Joint-EO methods are equal. First, we present conditions for the linear loss (2).

PROPOSITION 3.1. *Assume that the objective function $\mathbb{E}_{x|\theta}[f(\boldsymbol{x}, \boldsymbol{y})]$ 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 $\mathbb{E}_{x|\theta}[f(\boldsymbol{x}, \boldsymbol{y})]$, we have that x_i and x_j, as well as θ_i and θ_j are independent. Then, the Bayes solution estimator obtained from the Joint-EO scheme under the loss (2) is equal to the one obtained from the Separate-EO scheme, i.e., $\hat{\boldsymbol{y}}^{J,L}(\bar{\boldsymbol{x}}) = \hat{\boldsymbol{y}}^S(\bar{\boldsymbol{x}})$.*

Next, we present sufficient conditions under which the Separate-EO and the Joint-EO for the loss (7) yield the same solution estimators.

PROPOSITION 3.2. *Assume that (1) has a unique optimal solution $\boldsymbol{y}^*(\boldsymbol{\theta})$ for any given $\boldsymbol{\theta}$. Assume also that each component of $\boldsymbol{y}^*(\boldsymbol{\theta})$ 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 x_i and x_j, as well as θ_i and θ_j are independent. Then, the Bayes estimator obtained from the Joint-EO scheme under the loss (7) is equal to the one obtained from the Separate-EO scheme, i.e., $\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}}) = \hat{\boldsymbol{y}}^S(\bar{\boldsymbol{x}})$.*

The result of Proposition 3.2 implies that when the optimal solution of (1) is unique and multilinear in the parameters, the Bayes solution estimator obtained from (10) is equal to the expected optimal solution $\mathbb{E}_{\theta|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\theta})]$ as it belongs to the feasible region even when it is not convex. This can be viewed as another sufficient condition for the Joint-EO solution estimator to be equal to the expected optimal solution, beside the convexity sufficient condition given in Corollary 2.6.

While the conditions of Propositions 3.1 and 3.2 are satisfied by many stochastic optimization problems, there are important classes of stochastic problems that do not satisfy these conditions. We next present three examples of a portfolio selection model in finance that illustrate situations where the three solution estimators introduced in Section 2 are equal and/or different.

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

Example 3.3. Consider a set of n assets whose returns, recorded in $\boldsymbol{x}$, are random and follow a distribution with joint density function $g(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ and $\Sigma \succ 0$ represent the mean vector and the covariance matrix of the asset returns, respectively. The goal is to find a portfolio selection strategy $\boldsymbol{y}$ that maximizes the total expected return against the risk. This problem can be modeled as (1), where $f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\intercal \boldsymbol{y} - \frac{\tau}{2}[(\boldsymbol{x}-\boldsymbol{\mu})^\intercal \boldsymbol{y}]^2$ for some nonnegative *risk aversion* factor τ. This reduces to the classical Markowitz (mean-variance) model ([25])

$$\max_{\boldsymbol{y} \in \mathcal{Y}} \ \boldsymbol{\mu}^\intercal \boldsymbol{y} - \frac{\tau}{2} \boldsymbol{y}^\intercal \Sigma \boldsymbol{y}. \tag{11}$$

where $\mathcal{Y} = \mathbb{R}^n$ under the assumptions that (i) both long and short portfolio are allowed and (ii) a riskless asset is available.

Assume now that $\boldsymbol{\mu}$ is unknown and has a prior distribution $\pi(\boldsymbol{\mu})$, while Σ is known. The unconstrained problem (11) is convex and has a unique optimal solution $\boldsymbol{y}^*(\boldsymbol{\mu}) = \frac{1}{\tau}\Sigma^{-1}\boldsymbol{\mu}$. Since the objective function of (11) is linear in $\boldsymbol{\mu}$, it follows from Proposition 3.1 that the solution estimators of the Separate-EO and the Joint-EO under the loss (2) are equal. Similarly, since $\boldsymbol{y}^*(\boldsymbol{\mu})$ is linear in $\boldsymbol{\mu}$, it follows from Proposition 3.2 that the solution estimators of the Separate-EO and the Joint-EO under the loss (7) are equal. Therefore, all three solution estimators are equal to $\hat{\boldsymbol{y}}^{J,L}(\bar{\boldsymbol{x}}) = \hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}}) = \hat{\boldsymbol{y}}^{S}(\bar{\boldsymbol{x}}) = \frac{1}{\tau}\Sigma^{-1}\hat{\boldsymbol{\mu}}^B(\bar{\boldsymbol{x}})$ where $\hat{\boldsymbol{\mu}}^B(\bar{\boldsymbol{x}})$ 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 (7).

Example 3.4. Consider the setting of Example 3.3. The goal is to find weights $\boldsymbol{y}$ that maximize the expected return of the portfolio subject to bounding the portfolio risk by a constant r^2. This problem can be modeled as (1), where $f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\intercal \boldsymbol{y}$, and $\mathcal{Y} = \{\boldsymbol{y} \in \mathbb{R}^n | \boldsymbol{y}^\intercal \Sigma \boldsymbol{y} \le r^2\}$. We obtain that

$$\max_{\boldsymbol{y} \in \mathcal{Y}} \ \boldsymbol{\mu}^\intercal \boldsymbol{y}. \tag{12}$$

Assume now that $\boldsymbol{\mu}$ is unknown and has a prior distribution $\pi(\boldsymbol{\mu})$. Since the objective function of (12) is linear in $\boldsymbol{\mu}$, it follows from Proposition 3.1 that the solution estimators of the Separate-EO and the Joint-EO under the loss (2) are equal, i.e., $\hat{\boldsymbol{y}}^{J,L}(\bar{\boldsymbol{x}}) = \hat{\boldsymbol{y}}^{S}(\bar{\boldsymbol{x}}) = \frac{\hat{\boldsymbol{\mu}}^B(\bar{\boldsymbol{x}})}{||\hat{\boldsymbol{\mu}}^B(\bar{\boldsymbol{x}})||}$. On the other hand, the unique optimal solution of (12) is obtained as $\boldsymbol{y}^*(\boldsymbol{\mu}) = \frac{\Sigma^{-1}\boldsymbol{\mu}}{||\Sigma^{-1/2}\boldsymbol{\mu}||} r$, which is not a linear function of $\boldsymbol{\mu}$. As a result, the conditions of Proposition 3.2 are not satisfied, and therefore the solution estimators of the Separate-EO and the Joint-EO under the loss (7) can be different. This yields $\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}}) = \mathbb{E}_{\mu|\bar{x}}\left[\frac{\Sigma^{-1}\boldsymbol{\mu}}{||\Sigma^{-1/2}\boldsymbol{\mu}||} r\right]$, a quantity that can be very difficult to compute depending on the distributions. ■

Finally, we illustrate the converse of Example 3.4, where the solution estimators of the Separate-EO and the Joint-EO under the loss (7) are equal, but different from that of the Joint-EO under the loss (2).

Example 3.5. Consider the setting of Example 3.3, where the mean and covariance matrix of the asset returns are known, but the risk aversion factor τ is random with exponential distribution $\tau \sim \mathrm{Exp}(\lambda)$. This problem can be modeled as (1), where $f(\boldsymbol{\tau}, \boldsymbol{y}) = \boldsymbol{\mu}^\intercal \boldsymbol{y} - \frac{\tau}{2} \boldsymbol{y}^\intercal \Sigma \boldsymbol{y}$, and $\mathcal{Y} = \mathbb{R}^n$. We obtain that

$$\max_{\boldsymbol{y} \in \mathcal{Y}} \ \boldsymbol{\mu}^\intercal \boldsymbol{y} - \frac{1}{2\lambda} \boldsymbol{y}^\intercal \Sigma \boldsymbol{y}. \tag{13}$$

Assume now that parameter λ has a prior $\pi(\lambda)$. The unique optimal solution of the problem is $\boldsymbol{y}^*(\lambda) = \lambda\Sigma^{-1}\boldsymbol{\mu}$. Since $\boldsymbol{y}^*(\lambda)$ is linear in λ, it follows from Proposition 3.2

that the solution estimators of the Separate-EO and the Joint-EO under the loss (7) are equal, i.e., $\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}}) = \hat{\boldsymbol{y}}^{S}(\bar{\boldsymbol{x}}) = \hat{\lambda}^B(\bar{x})\Sigma^{-1}\boldsymbol{\mu}$ where $\hat{\lambda}^B(\bar{x})$ is the Bayes estimator (posterior mean) of λ, which is a shrinkage vector under exponential distributions. On the other hand, since the objective function of (13) is not linear in λ, it does not satisfy the conditions of Proposition 3.1, and therefore the solution estimators of the Separate-EO and the Joint-EO under the loss (2) can be different. In particular, $\hat{\boldsymbol{y}}^{J,L}(\bar{\boldsymbol{x}})$ is the maximizer of $\max_{\boldsymbol{y}\in\mathcal{Y}} \ \boldsymbol{\mu}^\intercal \boldsymbol{y} - \mathbb{E}_{\lambda|\bar{x}}[\frac{1}{2\lambda}]\boldsymbol{y}^\intercal\Sigma\boldsymbol{y}$. This yields $\hat{\boldsymbol{y}}^{J,L}(\bar{\boldsymbol{x}}) = \frac{\Sigma^{-1}\boldsymbol{\mu}}{\mathbb{E}_{\lambda|\bar{x}}[\frac{1}{\lambda}]}$, which may be computable only numerically depending on the distributions. ■

3.2. Different separate and joint estimators. In this section, we address the question of how the risk of the Separate-EO method compares to that of the Joint-EO method when their estimators are different. The next two examples show that the Separate-EO solution can yield arbitrarily poor solutions for general problem structures under the linear loss (2) and the quadratic loss (7), respectively.

Example 3.6. Assume that random variables x_i for $i \in [n]$ are independent and follow a normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$ where μ_i is unknown and σ_i^2 is known. Assume also that the parameters μ_i for $i \in [n]$ follow a normal distribution $\mathcal{N}(\lambda_i, \delta_i^2)$ where λ_i and δ_i^2 are known. It follows from Appendix B.1 that the posterior $\mu_i|\bar{x}_i$ has a normal distribution $\mathcal{N}(\eta_i, \xi_i^2)$ where $\eta_i = \frac{\sigma_i^2}{\sigma_i^2+\delta_i^2}\lambda_i + \frac{\delta_i^2}{\sigma_i^2+\delta_i^2}\bar{x}_i$ and $\xi_i^2 = \frac{\sigma_i^2\delta_i^2}{\sigma_i^2+\delta_i^2}$. Consider an instance of stochastic program (1) where the objective function is $\mathbb{E}_{x|\mu}[f(\boldsymbol{x}, \boldsymbol{y})] = \sum_{i=1}^n \mu_i^2 y_i$ and $\mathcal{Y}$ is an n-simplex, *i.e.,*

$$\max \left\{ \sum_{i=1}^n \mu_i^2 y_i \,\middle|\, \sum_{i=1}^n y_i = 1, y_i \geq 0, \forall i \in [n] \right\}. \tag{14}$$

To obtain the Bayes solution estimator for the Joint-EO method under the linear loss (2), we solve (14) by computing the posterior expectation of its objective function, see (4). We obtain the objective function $\sum_{i=1}^n (\eta_i^2 + \xi_i^2) y_i$ since $\mathbb{E}_{\mu|\bar{x}}[\mu_i^2] = \eta_i^2 + \xi_i^2$. It is easy to verify that the optimal solution (Bayes solution estimator) is attained at $\boldsymbol{y}^{J,L}(\bar{\boldsymbol{x}}) = \boldsymbol{e}^j$ where j is the index of the maximum value among $\{\eta_i^2 + \xi_i^2\}_{i\in[n]}$ while the minimizer (worst solution) of the above problem is attained at $\boldsymbol{y} = \boldsymbol{e}^k$ where k is the index of the minimum value among $\{\eta_i^2 + \xi_i^2\}_{i\in[n]}$. Now we calculate the estimator obtained from the Separate-EO method. To this end, we compute the Bayes estimator of the unknown parameter $\boldsymbol{\mu}$ under the squared error loss and use it in (14). As discussed before, this estimator is the posterior mean $\boldsymbol{\eta}$ and therefore the resulting objective function is $\sum_{i=1}^n \eta_i^2 y_i$. Using similar arguments as above, we obtain that the optimal solution of this problem is $\boldsymbol{y}^S(\bar{\boldsymbol{x}}) = \boldsymbol{e}^l$ 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 $\bar{\boldsymbol{x}}$ 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 arbitrary weak compared to the Joint-EO estimator.

For a numerical illustration of this result, assume that $n = 2$, $\bar{x}_1 = 2$, $\bar{x}_2 = 1$, $\lambda_1 = \lambda_2 = 0$, $\sigma_1 = \sigma_2 = 1$, $\delta_1 = 1$ and $\delta_2 = 3$. We compute $\eta_1 = 1$, $\eta_2 = \frac{9}{10}$, $\xi_1^2 = \frac{1}{2}$ and $\xi_2^2 = \frac{9}{10}$. It follows that $\eta_1^2 > \eta_2^2$ and $\eta_1^2 + \xi_1^2 < \eta_2^2 + \xi_2^2$. This shows that the Bayes solution estimator for the Joint-EO is $\boldsymbol{y}^{J,L}(\bar{\boldsymbol{x}}) = (0, 1)$ and the solution estimator for the Separate-EO is $\boldsymbol{y}^S(\bar{\boldsymbol{x}}) = (1, 0)$. ■

Example 3.7. Assume the setting of Example 3.6. Consider an instance of stochastic program (1) where the objective function is $\mathbb{E}_{x|\mu}[f(\boldsymbol{x},\boldsymbol{y})] = \sum_{i=1}^{n} \mu_i y_i$ and $\mathcal{Y}$ is a unit-ball in $\mathbb{R}^n$, *i.e.,*

$$\max\left\{\sum_{i=1}^{n} \mu_i y_i \,\middle|\, \sum_{i=1}^{n} y_i^2 \leq 1\right\}. \tag{15}$$

For any $\boldsymbol{\mu} \in \mathbb{R}^n \setminus \{0\}$, the unique optimal solution of the problem is $\boldsymbol{y}^*(\boldsymbol{\mu}) = \frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}$. As a result, the Separate-EO method yields the solution estimator $\boldsymbol{y}^S(\bar{\boldsymbol{x}}) = \frac{\boldsymbol{\eta}}{||\boldsymbol{\eta}||}$. Now we calculate the estimator obtained from the Joint-EO method under the quadratic loss. Since the optimal solution of (15) is unique and its feasible region is convex, it follows from Corollary 2.6 that $\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}}) = \mathbb{E}_{\mu|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\mu})] = \mathbb{E}_{\mu|\bar{x}}\left[\frac{\boldsymbol{\mu}}{||\boldsymbol{\mu}||}\right]$. 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 $\boldsymbol{\mu}|\bar{\boldsymbol{x}}$. When this distribution is non-degenerate (can assume multiple distinct values), the expected vector $\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}})$ belongs to the interior of the unit-ball. On the other hand, the Separate-EO solution $\boldsymbol{y}^S(\bar{\boldsymbol{x}})$ 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 $\bar{\boldsymbol{x}}$ are such that $|\eta_i| = \left|\frac{\sigma_i^2}{\sigma_i^2+\delta_i^2}\lambda_i + \frac{\delta_i^2}{\sigma_i^2+\delta_i^2}\bar{x}_i\right| \leq \epsilon$ for all $i \in [n]$ and a sufficiently small but positive ϵ. Due to the symmetry of the feasible region and the posterior distribution of $\boldsymbol{\mu}|\bar{\boldsymbol{x}}$ 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 of this result, assume that $n = 2$, $\bar{x}_1 = 1.001$, $\bar{x}_2 = -1$, $\lambda_1 = -1, \lambda_2 = 1$, $\sigma_1 = \sigma_2 = \delta_1 = \delta_2 = 1$. We compute $\eta_1 = 0.0005$, $\eta_2 = 0$ and $\xi_1^2 = \xi_2^2 = \frac{1}{2}$. 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 $\boldsymbol{y} = (1, 0)$. ■

Examples 3.3–3.5 show that a combination of objective function, constraints and probability distributions can lead to different solution estimators in form and complexity. Examples 3.6 and 3.7 show that these factors can even affect the quality of the solution estimators. For instance, if the parameters of the distributions in Example 3.6 are such that the addition of the posterior variance ξ^2 does not change the ranking of the components of $\boldsymbol{\eta}$, the Separate-EO and the Joint-EO give the same solution. Similarly, if the constraint set of (15) in Example 3.7 is replaced with the unit-ball surface, i.e., $\sum_{i=1}^{n} y_i^2 = 1$, it can be verified from directional statistics that the Separate-EO and the Joint-EO always give the same solution.

Inspired by these results, we next formalize the roles of problem structures and distributions by providing a closer comparison between the solution estimators for certain optimization models.

4. Piecewise Linear Structures. As given in Proposition 3.1 (resp. Propositions 3.2), the solution estimators obtained from the Joint-EO and Separate-EO are equal when the objective function (resp. optimal solution) is multilinear in the unknown parameters under the suitable independence criteria. Many optimization problems satisfy these conditions and hence provide identical results for both estimation methods. In this section, we explore other problem structures that do not satisfy the multilinearity conditions. Our goal is to compare the risk quality and computational complexity of different solution estimators for several classes and distributions.

4.1. Piecewise Linear Objective Functions. In this section, we study a class of stochastic program (1) that has a univariate piecewise linear objective function with a single breakpoint. Define $f(x,y):\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ as

$$f(x,y)=\begin{cases}\bar{f}(x,y), & \text{if } x\le y\\ \tilde{f}(x,y), & \text{if } x\ge y\end{cases}, \tag{16}$$

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

The piecewise linear function as described in (16) is 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 x follows a distribution $g(x|\theta)$ with a random parameter θ that has a prior $\pi(\theta)$. Let y be the decision variable representing the order quantity. Then, the profit is $f(x,y)=s\min\{y,x\}-cy$ which is of the form (16) where $\bar{f}(x,y)=sx-cy$ and $\tilde{f}(x,y)=(s-c)y$.

Other areas of application for piecewise linear functions with uncertain breakpoint (knot) emerge in modeling two-phase linear-linear processes such as latent growth behaviors [19]. These models are widely used to describe developmental processes in education and psychology. For instance, the acquisition of foundational vocabulary knowledge may progress in two phases where the functional form in each phase may well be different. The time at which the trajectory for behavior transitions from one phase to the other (i.e., the knot) is important scientifically and often marks a substantive watershed moment (e.g., a level of proficiency has been attained) or suggests when an intervention may be most beneficial.

Assume that $\lim_{t\to-\infty} tG(t|\theta)=0$, a property that holds for most of the standard probability distributions. Then, using an integration by part, the expectation of the function (16) is computed as

$$\begin{aligned}\mathbb{E}_{x|\theta}[f(x,y)] &= \left[\int_{z\le y}\bar{f}(z,y)g(z|\theta)dz+\int_{z\ge y}\tilde{f}(z,y)g(z|\theta)dz\right]\\ &= \tilde{a}\mathbb{E}_{x|\theta}[x]+\tilde{b}y+\tilde{c}+(\tilde{a}-\bar{a})\int_{z\le y}G(z|\theta)dz.\end{aligned} \tag{17}$$

It follows from (17) that the function $\mathbb{E}_{x|\theta}[f(x,y)]$, (i) is infinitely differentiable in y, (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 $y^*(\theta)=G^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\middle|\theta\right)$ when it is not linear, i.e., $\bar{a}-\tilde{a}\neq 0$.

Next, we consider the univariate unconstrained stochastic program

$$\max_{y\in\mathbb{R}}\mathbb{E}_{x|\theta}[f(x,y)], \tag{18}$$

where $f(x,y)$ is a piecewise linear objective function as defined in (16), and where $\bar{a}-\tilde{a}>0$ so that (18) is convex. Assume that the random variable x follows distribution $g(x|\theta)$ and the parameter θ is random with a prior $\pi(\theta)$. Given an observation $\bar{x}$ from the distribution $g(\bar{x}|\theta)$, Corollary 4.1 presents the Separate-EO estimator $\hat{y}^S(\bar{x})$ and the Joint-EO estimators $\hat{y}^{J,L}(\bar{x})$ and $\hat{y}^{J,Q}(\bar{x})$ under the linear and quadratic loss. This result is obtained as a direct consequence of Proposition 2.4 and Corollary 2.6.

COROLLARY 4.1. *Consider the piecewise linear problem (18), and assume that an observation $\bar{x}$ from the distribution $g(\bar{x}|\theta)$ is available. Then,*

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

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

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

It is clear from the result of Corollary 4.1 that the solution estimators depend on the distribution of x and its corresponding prior. Below, we consider three common conjugate families of likelihood-prior pairs and discuss the complexity of obtaining the solutions estimators together with approximation results on their risk differences. We refer the reader to Appendix B for a review of relevant conjugate distributions and the corresponding posterior and predictive distributions.

4.1.1. Normal Likelihood with Normal Prior.

PROPOSITION 4.2. *Consider the stochastic problem (18). Assume that the likelihood distribution is normal with $x \sim N(\mu, \sigma^2)$ and the prior distribution is normal with $\mu \sim N(\mu_0, \delta^2)$. Assume further that the parameters σ^2, μ_0 and δ^2 are known, and an observation $\bar{x}$ is drawn from the likelihood distribution. Then, we have*

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

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

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 4.1(i), we have

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

where $G(x)$ is the cdf of a normal random variable with mean $\hat{\mu}^B(\bar{x}) := \rho\mu_0 + (1-\rho)\bar{x}$ and variance σ^2. Hence, we obtain $\hat{y}^S(\bar{x})$ as stated. We also observe that $\hat{y}^*(\mu)$ is linear in μ. Therefore, by Proposition 3.2, we conclude that $\hat{y}^S(\bar{x}) = \hat{y}^{J,Q}(\bar{x})$.

(ii): Due to Corollary 4.1(ii), we have

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

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

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 discuss the risk performance guarantee between the two solution estimators by computing the difference in their risk values. To this end, we first give some properties of a particular function that will be used later to derive the results.

Remark 4.3. 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 4.4. *Let $2\tilde{b} \geq \bar{a} - \tilde{a}$. Under the assumptions of Proposition 4.2,*

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

Proof. We first compute an upper bound on the difference in the loss values of the estimators. Define $F(y) := \mathbb{E}_{x|\mu}[f(x,y)]$ as the objective function of (18). Then,

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

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 $F(y)$ at point $\hat{y}^{J,L}(\bar{x})$ about $\hat{y}^S(\bar{x})$, the second inequality follows from (i) $\hat{y}^{J,L}(\bar{x}) - \hat{y}^S(\bar{x}) = \Phi^{-1}\left(\frac{\tilde{b}}{\bar{a}-\tilde{a}}\right)\sigma(\sqrt{2-\rho}-1)$ because of Proposition 4.2, (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) $F'(y) \leq \tilde{b}$ for all $y \in \mathbb{R}$ which is deduced from (17), the second equality holds due to the definition of $d(\kappa)$ with $\kappa := \sigma/\delta$ given in Remark 4.3 and $\rho = \frac{\sigma^2}{\sigma^2+\delta^2}$, and the last inequality follows from Remark 4.3(iii). Since the difference in the loss of the two estimators in the above expression is independent of both μ and $\bar{x}$, the risk difference obtained by taking the expectations $\mathbb{E}_\mu \mathbb{E}_{\bar{x}|\mu}$ as given in (3) remains unchanged, yielding the result. ∎

The following results is obtained similarly to that of Proposition 4.4.

PROPOSITION 4.5. *Let $2\tilde{b} \leq \bar{a} - \tilde{a}$. Under the assumptions of Proposition 4.2,*

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

4.1.2. Exponential Likelihood with Gamma Prior.

PROPOSITION 4.6. *Consider the stochastic problem (18). Assume that the likelihood distribution is exponential with $x \sim Exp(\lambda)$ and the prior distribution is gamma with $\lambda \sim Gamma(\alpha, \beta)$. Assume further that the shape and rate hyperparameters α and β are known, and a realization $\bar{x}$ is observed. Then,*

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

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

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

Proof. (i): Due to Corollary 4.1(i), we have

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

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

(ii): Due to Corollary 4.1(ii), we have

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

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

(iii): We write that

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

where the first equality follows from Corollaries 2.6 and 4.1(iii), the second equality holds since the posterior distribution $\Pi(\lambda|\bar{x})$ is gamma with the shape and rate parameters $\alpha+1$ and $\beta+\bar{x}$ 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 $\hat{y}^S(\bar{x})$ and $\hat{y}^{J,L}(\bar{x})$ as well as $\hat{y}^{J,Q}(\bar{x})$.

PROPOSITION 4.7. *Let $\alpha > 1$. Under the assumptions of Proposition 4.6,*

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

Proof. We first compute an upper bound on the difference in the loss values of the estimators. Define $F(y) := \mathbb{E}_{x|\mu}[f(x,y)]$ as the objective function of (18).

We first observe that $\hat{y}^{J,L} > \hat{y}^S$ due to the relationship

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

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 $F(y)$ at point $\hat{y}^{J,L}(\bar{x})$ about $\hat{y}^S(\bar{x})$, the second inequality follows from (i) $\hat{y}^{J,L}(\bar{x}) - \hat{y}^S(\bar{x}) = (\beta+\bar{x})\left[\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)^{1/(\alpha+1)} - 1\right] - \frac{\beta+\bar{x}}{\alpha+1}\ln\left(\frac{\bar{a}-\tilde{a}}{\bar{a}-\tilde{a}-\tilde{b}}\right)$ due to Proposition 4.6, (ii)

$\hat{y}^{J,L}(\bar{x}) > \hat{y}^S(\bar{x})$ since $\kappa^{1/a} - 1 > \frac{\ln(\kappa)}{a}$ for all $\kappa, a > 1$, and (iii) $F'(y) \leq \tilde{b}$ for all $y \in \mathbb{R}$ which is deduced from (17). To obtain the risk difference, we take the expectations $\mathbb{E}_\lambda \mathbb{E}_{\bar{x}|\lambda}$ from the last term, which yields the desired result using $\mathbb{E}_\lambda \mathbb{E}_{\bar{x}|\lambda}[\bar{x}] = \frac{\beta}{\alpha-1}$. □

PROPOSITION 4.8. *Let $\alpha > 2$. Under the assumptions of Proposition 4.6,*

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

Proof. We write that

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

where the first equality follows from Proposition 4.6 and the definition (7) of the risk under quadratic loss, and the second equality holds because $\mathbb{E}_{x|\lambda}[x] = \frac{1}{\lambda}$ and $\mathbb{E}_{x|\lambda}[x^2] = \frac{2}{\lambda^2}$. Using similar arguments, we can compute

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

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

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

from which the result follows. □

4.1.3. Geometric Likelihood with Beta Prior.

PROPOSITION 4.9. *Consider the stochastic problem (18). Assume that the likelihood distribution is geometric with $x \sim Geo(p)$ and the prior distribution is beta with $p \sim Beta(\alpha, \beta)$. Assume further that the shape parameters α and β are known, and a realization $\bar{x}$ is observed from the likelihood. Then,*

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

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

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

Proof. (i): Due to Corollary 4.1(i), we have

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

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

(ii): Due to Corollary 4.1(ii), we have

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

where H is the cdf of posterior predictive distribution. The results follows due to the relationship between the cdf H and its pmf h given in (30) in Appendix B.

(iii): We write that

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

where the first relation follows from Corollaries 2.6 and 4.1(iii) and the second relation holds since the posterior distribution $\Pi(p|\bar{x})$ is beta with parameters $\alpha+1$ and $\beta+\bar{x}-1$. ∎

We note that $\hat{y}^S(\bar{x})$ can be approximated by a closed form expression while the approximations of $\hat{y}^{J,L}(\bar{x})$ and $\hat{y}^{J,Q}(\bar{x})$ 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.

In Table 1, we summarize the conclusions drawn for the stochastic problem (18).

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

TABLE 1

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.

4.2. Sum of Piecewise Linear Functions. In this section, we study the stochastic program

$$\max_{y\in\mathbb{R}^n} \mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta}}\left[\sum_{i=1}^n f_i(x_i,y)\right], \tag{19}$$

where $f_i(x_i,y)$ is a piecewise linear objective function of the form (16), i.e., $\bar{f}_i(x_i,y) = \bar{a}_i x_i + \bar{b}_i y + \bar{c}_i$ for $x_i \leq y$, and $\tilde{f}_i(x_i,y) = \tilde{a}_i x_i + \tilde{b}_i y + \tilde{c}_i$ for $x_i \geq y$. We assume that $\bar{a}_i - \tilde{a}_i > 0$ for each $i \in [n]$, so that (19) is convex. In this problem, y is the single decision variable and x_i is a random variable with probability distribution $g_i(x_i|\theta_i)$ and the random parameter θ_i has prior distribution $\pi_i(\theta_i)$. In this setting, variables x_i are not required to be independent. We further assume that an observation $\bar{x}_i$ drawn from $g_i(\bar{x}_i|\theta_i)$ for $i \in [n]$ is available.

Stochastic problems of the form (19) have applications in median and location-allocation problems in facility planning. For instance, consider the one-dimensional stochastic median problem, where the position x_i of n points on the line is random and each comes from a distribution $g_i(x_i|\boldsymbol{\theta}_i)$ for $i \in [n]$. Let y be the decision variable representing the location of the median. The goal is to find y that minimizes the total distance from the random points, i.e., $\sum_{i=1}^{n} |x_i - y|$. This problem can be formulated as (19) where $f_i(x_i, y) = |x_i - y|$, $\bar{f}_i(x_i, y) = y - x_i$ and $\tilde{f}_i(x_i, y) = x_i - y$.

Using (17), the objective function of (19) can be computed as

$$\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta}}[\sum_{i=1}^{n} f_i(x_i, y)] = \sum_{i=1}^{n} \left[\tilde{a}_i \mathbb{E}_{x_i|\theta_i}[x_i] + \tilde{b}_i y + \tilde{c}_i + (\tilde{a}_i - \bar{a}_i) \int_{z_i \le y} G_i(z_i|\theta_i) dz_i \right], \tag{20}$$

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 following equation

$$\sum_{i=1}^{n} (\tilde{a}_i - \bar{a}_i) G_i(y|\theta_i) + \sum_{i=1}^{n} \tilde{b}_i = 0. \tag{21}$$

We next give the Separate-EO estimator $\hat{y}^S(\bar{x})$ and the Joint-EO estimator $\hat{y}^{J,L}(\bar{x})$ via univariate root finding using (21). We note that the Joint-EO estimator $\hat{y}^{J,Q}(\bar{x})$ is not presented since the explicit solution for (21) is not computable in general.

PROPOSITION 4.10. *Consider problem* (19). *Assume that, for each $i \in [n]$, the likelihood distribution has the pdf $g_i(x_i|\theta_i)$ and the prior distribution has the pdf $\pi_i(\theta_i)$. Assume further that an observation $\bar{x}_i$ is available for $i \in [n]$. Then,*

(i) $\hat{y}^S(\bar{y})$ is the solution of

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

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

(ii) $\hat{y}^{J,L}$ is the solution of

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

where $H_i(x)$ is the cdf of the posterior predictive distribution of x_i given observation $\bar{x}_i$.

Next, we give a generic method to compute an upper bound on the risk difference between the Separate-EO estimator $\hat{y}^S(\bar{x})$ and the Joint-EO estimator $\hat{y}^{J,L}(\bar{x})$. This bound is obtained in terms of the individual Separate and Joint solution estimators for each random variable x_i independently, as studied in Section 4.1.

PROPOSITION 4.11. *Consider the setting in Proposition 4.10. Assume that the likelihood $g_i(x_i|\theta_i)$ and the posterior predictive $h_i(x_i|\bar{x}_i)$ are positive at all points in* *the domain. Then,*

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

where the expectation $\mathbb{E}_{\bar{\boldsymbol{x}}}$ is taken with respect to the marginal distribution of $\bar{x}$, $K=\max\{|\sum_{i=1}^n\tilde{b}_i|,|\sum_{i=1}^n\tilde{b}_i+\tilde{a}_i-\bar{a}_i|\}$, $\hat{y}_i^S(\bar{x}_i)$ and $\hat{y}_i^{J,L}(\bar{x}_i)$ are the Separate-EO and Joint-EO estimators if there was only a single random variable x_i, as computed in Corollary 4.1.

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

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

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 $\hat{y}^S(\bar{x})\in\left[\min_i\left\{G_i^{-1}\left(\frac{\tilde{b}_i}{\bar{a}_i-\tilde{a}_i}\middle|\hat{\theta}_i^B(\bar{x}_i)\right)\right\},\max_i\left\{G_i^{-1}\left(\frac{\tilde{b}_i}{\bar{a}_i-\tilde{a}_i}\middle|\hat{\theta}_i^B(\bar{x}_i)\right)\right\}\right]$. A similar argument shows that $\hat{y}^{J,L}(\bar{x})\in\left[\min_i\left\{H_i^{-1}\left(\frac{\tilde{b}_i}{\bar{a}_i-\tilde{a}_i}\middle|\bar{x}_i\right)\right\},\max_i\left\{H_i^{-1}\left(\frac{\tilde{b}_i}{\bar{a}_i-\tilde{a}_i}\middle|\bar{x}_i\right)\right\}\right]$.

We now give an upper bound on the loss values of these estimators. Let $F(y)=\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta}}\left[\sum_{i=1}^n f_i(x_i,y)\right]$ as the objective function of (19). Since F is concave, we have

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

where the last inequality holds because (i) $|F'(\hat{y}^S)|\leq K$ as $F'(\hat{y}^S)\in[\sum_{i=1}^n\tilde{b}_i+\tilde{a}_i-\bar{a}_i,\sum_{i=1}^n\tilde{b}_i]$ due to (20), and (ii) the previously derived intervals for $\hat{y}^S(\bar{x})$ and $\hat{y}^{J,L}(\bar{x})$, 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 $\mathbb{E}_\theta\mathbb{E}_{\bar{x}|\theta}$ or equivalently $\mathbb{E}_{\bar{x}}\mathbb{E}_{\theta|\bar{x}}$ of both sides of the above expression. Since this expression is independent of θ the desired result is obtained by only taking the expectation $\mathbb{E}_{\bar{x}}$. □

The result of Proposition 4.11 requires an additional expectation with respect to the marginal distribution of $\bar{x}$, over the maximum of random variables. The following result can be useful to obtain closed-form bounds for such quantities.

PROPOSITION 4.12 ([1]). *Let $z_1,\ldots,z_n$ be random variables. Then,*

(i) $\mathbb{E}[\max_i\{z_i\}]\leq\max_i\{\mathbb{E}[z_i]\}+\sqrt{\frac{n-1}{n}\sum_{i=1}^n \mathit{Var}(z_i)}$.

(ii) $\mathbb{E}[\min_i\{z_i\}]\geq\min_i\{\mathbb{E}[z_i]\}-\sqrt{\frac{n-1}{n}\sum_{i=1}^n \mathit{Var}(z_i)}$.

The bound given in Proposition 4.11 can be hard to compute in closed-form for general classes of the stochastic program (19). However, it can be computed explicitly for certain objective functions and probability distributions. We next present two instances of a certain class. In particular, we consider the one-dimensional stochastic median problem where $f_i(x_i,y)=|x_i-y|$ for $i\in[n]$, and compute the bound on the risk of the Separate-EO and Joint-EO solution estimators for two conjugate pairs.

4.2.1. Normal Likelihood with Normal Prior.

PROPOSITION 4.13. *Consider the one-dimensional stochastic median problem. Assume that, for each $i \in [n]$, the likelihood distribution is normal with $x_i \sim N(\mu_i, \sigma_i^2)$ and the prior distribution is normal with $\mu_i \sim N(\mu_i^0, \delta_i^2)$. Assume further that the parameters σ_i^2, μ_i^0, δ_i^2 are known, and a realization of locations $\bar{x}_i$ is observed. Then,*

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

Proof. Proposition 4.2 implies that $\nu_i := \hat{y}_i^S(\bar{x}_i) = \hat{y}_i^{J,L}(\bar{x}_i) = \rho_i\mu_i^0 + (1-\rho_i)\bar{x}_i$ 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 4.11 to obtain

$$\mathcal{R}^L\left(y^*(\boldsymbol{\mu}), \hat{y}^S(\bar{x})\right) - \mathcal{R}^L\left(y^*(\boldsymbol{\mu}), \hat{y}^{J,L}(\bar{x})\right) \leq n\mathbb{E}_{\bar{\boldsymbol{x}}}[\max_i\{\nu_i\} - \min_i\{\nu_i\}]. \tag{22}$$

Now, we compute an upper bound on the right-hand-side using Proposition 4.12. Since $\bar{x}_i \sim \mathrm{N}(\mu_i^0, \sigma_i^2 + \delta_i^2)$ 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 result of Proposition 4.12, we obtain

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

Plugging in these bounds for into (22) gives the desired conclusion. ∎

4.2.2. Exponential Likelihood with Gamma Prior.

PROPOSITION 4.14. *Consider the one-dimensional stochastic median problem. Assume that, for each $i \in [n]$, the likelihood distribution is exponential with $x_i \sim Exp(\lambda_i)$ and the prior distribution is gamma with $\lambda_i \sim Gamma(\alpha_i, \beta_i)$. Assume further that the parameters α_i and β_i are known with $\alpha_i > 2$, and that a realization of locations $\bar{x}_i$ is observed for $i \in [n]$. Then, we have*

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

Proof. Proposition 4.6 implies that $\hat{y}_i^S(\bar{x}_i) = \frac{\beta_i+\bar{x}_i}{\alpha_i+1}\ln 2$ and $\hat{y}_i^{J,L}(\bar{x}_i) = (\beta_i + \bar{x}_i)(\sqrt[\alpha_i+1]{2}-1)$ 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 $\hat{y}_i^S(\bar{x}_i) < \hat{y}_i^{J,L}(\bar{x}_i)$ for each $i \in [n]$. Applying Proposition 4.11, we obtain

$$\mathcal{R}^L\left(y^*(\boldsymbol{\lambda}), \hat{y}^S(\bar{x})\right) - \mathcal{R}^L\left(y^*(\boldsymbol{\lambda}), \hat{y}^{J,L}(\bar{x})\right) \leq n\mathbb{E}_{\bar{\boldsymbol{x}}}[\max_i\{\hat{y}_i^{J,L}(\bar{x}_i)\} - \min_i\{\hat{y}_i^S(\bar{x}_i)\}]. \tag{23}$$

Now, we compute an upper bound on the right-hand-side using Proposition 4.12. Since $\bar{x}_i \sim \mathrm{Lomax}(\beta_i, \alpha_i)$ for $i \in [n]$, we have

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

Since both $\hat{y}_i^S(\bar{x}_i)$ and $\hat{y}_i^{J,L}(\bar{x}_i)$ are affine transformations of x_i, we can easily obtain their mean and variance as well. Finally, plugging in the resulting bounds for $\mathbb{E}[\min_i\{\hat{y}_i^S(\bar{x}_i)\}]$ and $\mathbb{E}[\max_i\{\hat{y}_i^{J,L}(\bar{x}_i)\}]$ into (23) gives the desired result. □

We conclude this section by a remark on a generalization of univariate stochastic programs with a piecewise linear objective function to a certain multivariate case where the previously derived results can apply.

Remark 4.15. Stochastic problem (19) can be generalized in a straightforward manner to the case where there are several decision variables as follows. Consider the objective function $\sum_{j=1}^m \sum_{i=1}^n f_i^j(x_i, y^j)$ where $f_i^j(x_i, y^j)$ is equal to $\bar{f}_i^j(x_i, y^j) = \bar{a}_i^j x_i + \bar{b}_i^j y^j + \bar{c}_i^j$ for $x_i \leq y^j$ and equal to $\tilde{f}_i^j(x_i, y^j) = \tilde{a}_i^j x_i + \tilde{b}_i^j y^j + \tilde{c}_i^j$ for $x_i \geq y^j$. Under the assumption that each x_i has a distinct distribution $g_i(x_i|\theta_i)$ and its parameter θ_i has a distinct prior $\pi_i(\theta_i)$, it is easy to verify that the objective function can be separated as a summation of univariate objective functions appearing in (19). Since the feasible region is unrestricted, the resulting problem reduces to m individual problems of the form (19) and therefore the results of Section 4.2 applies accordingly. This generalization models an m-dimensional stochastic median problem that has extensive applications in facility planning and assignment problems.

5. Geometric Structures. Another class of stochastic programs where the random element in the objective function does not appear in a multilinear form is the stochastic geometric programs (GP). GPs have applications in a wide variety of problems including circuit design, optimal control, nonlinear network design and chemical equilibrium problems; see [7] for a tutorial on geometric programming and for a full list of applications. Due to developments in designing effective algorithms for convex programs, the state-of-the-art software packages can solve large-size GPs efficiently for practical applications. For this reason, the tendency to model complicated problems as GPs either exactly or approximately has rapidly increased in the past decade [11]. This rise in the application of GPs has also lead to incorporating the stochastic nature of real-world problems into these models.

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 $z_i = e^{y_i}$ for all i and replace $f(z)$ with $\log f(e^y)$ in the model. Such problems can be formulated as

$$\min_{\boldsymbol{y} \in \mathcal{Y}} \left\{ \sum_k c_k e^{\boldsymbol{\alpha}_k^\intercal \boldsymbol{y}} \,\middle|\, \log f_j(e^y) \leq 0, \forall j = 1, \dots, m \right\}, \tag{24}$$

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

In a stochastic variant of (24), we assume that the exponents α_{ik} are random with distribution $g_{ik}(x_{ik}|\theta_{ik})$. Including all constraints in $\mathcal{Y}$, we write the general form of the stochastic geometric program as

$$\min_{\boldsymbol{y} \in \mathcal{Y}} \mathbb{E}_{x|\theta} \left[\sum_k c_k e^{\boldsymbol{x}_k^\intercal \boldsymbol{y}} \right]. \tag{25}$$

Under the assumption that the random exponents x_{ik} are independent, the objective function of (25) decomposes into separable terms as $\mathbb{E}_{x|\theta}\left[\sum_k c_k e^{\boldsymbol{x}_k^\intercal \boldsymbol{y}}\right] = \sum_k c_k \prod_i \mathbb{E}_{x_{ik}|\theta_{ik}}[e^{x_{ik} y_i}]$. The unique property of such decomposition is that the term $\mathbb{E}_{x_{ik}|\theta_{ik}}[e^{x_{ik}y_i}]$ is equal to the *moment generating function* $M_{x_{ik}|\theta_{ik}}(y_i)$ of the distribution $g_{ik}(x_{ik}|\theta_{ik})$. As a result, the stochastic geometric problem (25) reduces to

$$\min_{\boldsymbol{y}\in\mathcal{Y}} \sum_k c_k \prod_i M_{x_{ik}|\theta_{ik}}(y_i). \tag{26}$$

Assume now that the parameter θ_{ik} is random with prior $\pi(\theta_{ik})$. Then, using the results of Section 2, the Separate-EO and Joint-EO solution estimators for (26) under the linear loss (2) are obtained as follows.

COROLLARY 5.1. *Consider the stochastic geometric program* (26)*. Assume that* x_{ik} *has distribution* $g_{ik}(x_{ik}|\theta_{ik})$ *and its parameter* θ_{ik} *has prior* $\pi(\theta_{ik})$*. Further, assume that an observation* $\bar{x}_{ik}$ *is available. Then,*

(i) $y^S(\bar{x}) = \operatorname{argmin}_{\boldsymbol{y}\in\mathcal{Y}} \sum_k c_k \prod_i M_{x_{ik}|\hat{\theta}^B_{ik}}(y_i)$*, where* $\hat{\theta}^B_{ik}$ *is the Bayes estimator of* θ_{ik} *under the squared error loss.*

(ii) $y^{J,L}(\bar{x}) = \operatorname{argmin}_{\boldsymbol{y}\in\mathcal{Y}} \sum_k c_k \prod_i M_{x_{ik}|\bar{x}_{ik}}(y_i)$*, where* $M_{x_{ik}|\bar{x}_{ik}}$ *is the moment generating function of the posterior predictive distribution* $h(x_{ik}|\bar{x}_{ik})$*.*

Even though the result of Corollary 5.1 implies that both the Separate-EO and Joint-EO estimators can be obtained by solving a product of moment generating functions, their computational complexity can be different. In particular, for most of the standard likelihood functions, the moment generating function is readily computed in closed-form which makes the problem of obtaining the Separate-EO deterministic. On the other hand, for the posterior predictive distributions, the moment generating functions are hard to compute in closed-form even when the distributions belong to the conjugate family. For instance, for the exponential-gamma conjugate pair, it is known that the moment generating function of the posterior predictive distribution, Lomax, is mathematically intractable. This makes the computation of the Joint-EO estimator much harder, as it requires the employment of numerical algorithms to approximate the optimal solution. This class of stochastic programs contains problems where the trade-off between the computational efficiency and the risk quality of the Separate-EO and Joint-EO estimators becomes more noticeable. Below, we present a preliminary computation study, which illustrates this intuition.

Consider the stochastic geometric program (26) for the exponential-gamma conjugate pair with prior hyperparameters α_k and β_k. We assume that a set of observations $\bar{x}_k$, $k \in [K]$, is available. According to Corollary 5.1, obtaining the Separate-EO solution estimator amounts to solving the convex program

$$z^S := \min_{\boldsymbol{y}\in\mathcal{Y}} \left\{ \sum_{k=1}^K c_k \frac{\lambda_k(\bar{x}_k)}{\lambda_k(\bar{x}_k) - y_k} \right\}, \tag{27}$$

where $\lambda_k(\bar{x}_k) = \frac{\alpha_k+1}{\beta_k+\bar{x}_k}$. Here, we implicitly use the fact that the moment generating function of the exponential distribution is available in closed-form. However, this is not possible for the Lomax distribution, thereby making the use of sampling-based methods (such as SAA) necessary to obtain the Joint-EO solution estimator. In particular, let $\tilde{x}_{rk}$ be a Lomax random variable[1] with parameters $\beta_k + \bar{x}_k$ and $\alpha_k + 1$, for $k \in [K]$ and $r \in [R]$, where R denotes the sample size. Then applying the SAA method to (24) for the Lomax distribution, we can approximate the Joint-EO solution

[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

estimator with the solution of the following convex program:

$$z^{J,L}(R) := \min_{\boldsymbol{y}\in\mathcal{Y}} \left\{ \sum_{k=1}^{K} c_k \frac{1}{R} \sum_{r=1}^{R} e^{\tilde{x}_{rk} y_k} \right\}. \tag{28}$$

We will now discuss an experimental setting. We define the feasible region $\mathcal{Y}$ as the intersection of the standard simplex $\Delta_K := \{\boldsymbol{y} \in \mathbb{R}^K_+ : \sum_{k=1}^K y_k = 1\}$ with a polyhedron $\mathcal{P} := \{\boldsymbol{y} \in \mathbb{R}^K : A\boldsymbol{y} = \boldsymbol{b}\}$ 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 $f_j(e^y) = ce^{\boldsymbol{\alpha}^\intercal \boldsymbol{y}}$, and hence the log constraint as presented in (24) becomes linear in $\boldsymbol{y}$, i.e, $\boldsymbol{\alpha}^\intercal\boldsymbol{y} \leq -\log c$. We randomly generate the parameters according to the following rules:

$$\begin{array}{lll} c_k = 1 & \alpha_k \sim \text{Unif}(5,10) & \beta_k \sim \text{Unif}(5,10) \\ A_{mk} \sim \text{Unif}(1,10) & b_m \sim \text{Unif}(10,50) & \bar{x}_k = \dfrac{\alpha_k}{10(\beta_k-1)} \end{array}$$

In Table 2, we compare the Separate-EO (represented in columns with label SEO) and the Joint-EO (represented in columns with label JEO(R)) solution estimators with R = 100, 1000 and 10000 for five randomly generated instances in dimension $K = 1000$ with $M = 10$ linear constraints. The common characteristic of this family of instances is that the objective values $z^{J,L}(R)$ for the Joint-EO method exhibit a slow convergence behavior. 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. We also observe that obtaining the JEO(R) solution estimators becomes increasingly demanding with larger R values in terms of the computational effort. For example, for a moderate-sized instance considered here, it typically takes more than 2 minutes to obtain a solution estimator when a sample size of 10000 is used. For this problem instance, we ran into memory issues with 100000 replications. 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.

	SEO		JEO (100)		JEO (1000)		JEO (10000)	
ins.	obj. val.	time	obj. val.	time	obj. val.	time	obj. val.	time
1	1005.17	0.51	1012.74	1.12	1018.61	13.19	1025.29	140.77
2	1007.61	0.32	1009.93	1.25	1014.52	11.50	1019.54	145.72
3	1009.79	0.34	1012.65	1.02	1020.80	12.93	1047.55	102.43
4	1009.87	0.47	1027.26	1.20	1041.79	10.95	1097.27	169.69
5	1032.39	0.40	1018.28	1.47	1038.84	9.69	1068.64	135.85

TABLE 2

Computational results for randomly generated stochastic geometric programs with exponential-gamma pair ($K = 1000$, $M = 5$). Five problem instances (ins.) are solved, and the corresponding objective values (obj. val.) and computational times in seconds are recorded.

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, [5] argues that despite 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 suboptimal, enjoy a fast and stable solution process. This trade-off between the two schemes should be considered carefully when modeling stochastic problems of certain classes such as geometric programs.

6. Conclusion.

In the presence of uncertainty in optimization, random data is used to estimate the unknown elements of the underlying stochastic programs. Several techniques can be employed to find such estimators. Under a Bayesian framework, two of the most common estimation rules solve the estimation and optimization problems either separately or jointly. While it is intuitively obvious that the joint scheme provides better results in quality, analytical and computational studies on the trade-offs between the quality and complexity of these schemes are lacking. In this paper, we explore this direction to provide insight on the advantages and disadvantages of these two schemes. 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 two popular classes of nonlinear stochastic programs with piecewise linear and geometric structure. We give theoretical and computational evidence for an in-depth comparison between the risk value and complexity of different schemes for these problem classes.

Appendix A. Proofs.

Proof of Proposition 2.1. We write that

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

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

Proof of Proposition 2.4. Using the result of Proposition 2.1, it suffices to show that the objective function of (4) matches that of (5). We show the result for continuous distributions, as that of the discrete distribution follows similarly. Under the assumption that the objective function $f(\boldsymbol{x}, \boldsymbol{y})$ and all distribution of variables admit the interchange of integration order, we have that

$$\mathbb{E}_{\theta|\bar{x}}\mathbb{E}_{x|\theta}[f(\boldsymbol{x},\boldsymbol{y})] = \int_{\boldsymbol{\theta}}\int_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{y})g(\boldsymbol{x}|\boldsymbol{\theta})\Pi(\boldsymbol{\theta}|\bar{\boldsymbol{x}})d\boldsymbol{x}d\boldsymbol{\theta}$$

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

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 2.3. □

Proof of Proposition 2.5. We have that $\mathcal{R}^Q(\boldsymbol{y}^*(\boldsymbol{\theta}),\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}})) = \mathbb{E}_{\theta}\mathbb{E}_{\bar{x}|\theta}[\mathcal{L}^Q(\boldsymbol{y}^*(\boldsymbol{\theta}),\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}}))] = \mathbb{E}_{\bar{x}}\mathbb{E}_{\theta|\bar{x}}[\mathcal{L}^Q(\boldsymbol{y}^*(\boldsymbol{\theta}),\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}}))]$, where the first equality holds because of (8), and the second equality follows from changing the order of sequential expectations. According to the definition of Bayes estimator, we seek among all $\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}}) \in \mathcal{Y}$ an estimator that minimizes the risk $\mathcal{R}^Q(\boldsymbol{y}^*(\boldsymbol{\theta}),\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}}))$. We claim that any minimizer $\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}})$ of $\min_{\boldsymbol{y}\in\mathcal{Y}} \mathbb{E}_{\theta|\bar{x}}\left[\mathcal{L}^Q(\boldsymbol{y}^*(\boldsymbol{\theta}),\boldsymbol{y})\right]$ is also a minimizer of the risk $\mathcal{R}^Q(\boldsymbol{y}^*(\boldsymbol{\theta}),\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}}))$. To prove this claim, consider any estimator $\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}}) \neq \hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}})$. It follows from the assumption that $\mathbb{E}_{\theta|\bar{x}}[\mathcal{L}^Q(\boldsymbol{y}^*(\boldsymbol{\theta}),\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}}))] \leq \mathbb{E}_{\theta|\bar{x}}[\mathcal{L}^Q(\boldsymbol{y}^*(\boldsymbol{\theta}),\hat{\boldsymbol{y}}(\bar{\boldsymbol{x}}))]$. Taking the expectation $\mathbb{E}_{\bar{x}}[.]$ with respect to the marginal distribution of $\bar{\boldsymbol{x}}$ from both sides we obtain the desired result due to the chain relation in the first line. □

Proof of Proposition 3.1. It follows from the assumptions that $\mathbb{E}_{x|\theta}[f(\boldsymbol{x},\boldsymbol{y})]$ can be written as $\sum_{k=1}^{K} h_k(\boldsymbol{y})g_k(\boldsymbol{\theta})$ where $h_k(\boldsymbol{y}) : \mathbb{R}^m \to \mathbb{R}$ 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 $k \in \{1,\dots,K\}$, all variables x_l for $l \in I_k$ are independent, and so are all variables θ_l for $l \in I_k$. We compute the objective of the Joint-EO method as given in (4)

$$\mathbb{E}_{\theta|\bar{x}}\mathbb{E}_{x|\theta}[f(\boldsymbol{x},\boldsymbol{y})] = \sum_{k=1}^{K} h_k(\boldsymbol{y})\mathbb{E}_{\theta|\bar{x}}[g_k(\boldsymbol{\theta})] = \sum_{k=1}^{K} h_k(\boldsymbol{y})\prod_{i\in I_k}\mathbb{E}_{\theta_i|\bar{x}_i}[\theta_i], \tag{29}$$

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 θ_i is replaced by its posterior mean $\mathbb{E}_{\theta_i|\bar{x}_i}[\theta_i]$ in (1). The resulting objective function is (29). This shows that the objective of the Joint-EO and Separate-EO methods are equal. Since both have the same constraint set $\mathcal{Y}$, their optimal solutions are the same. □

Proof of Proposition 3.2. Under the assumption that $\boldsymbol{y}^*(\boldsymbol{\theta})$ is unique for any given $\boldsymbol{\theta}$, the Separate-EO solution estimator can be expressed as $\hat{\boldsymbol{y}}^S(\bar{\boldsymbol{x}}) = \boldsymbol{y}^*(\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{x}}))$ where $\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{x}})$ 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 $\hat{\boldsymbol{\theta}}^B(\bar{\boldsymbol{x}}) = \mathbb{E}_{\theta|\bar{x}}[\boldsymbol{\theta}]$. Therefore, the Separate-EO method yields $\hat{\boldsymbol{y}}^S(\bar{\boldsymbol{x}}) = \boldsymbol{y}^*(\mathbb{E}_{\theta|\bar{x}}[\boldsymbol{\theta}])$. We obtain that $\mathbb{E}_{\theta|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\theta})] = \boldsymbol{y}^*(\mathbb{E}_{\theta|\bar{x}}[\boldsymbol{\theta}])$ because of the linearity of the expectation operator and because of the independence of variables appearing in the products. As a result, $\mathbb{E}_{\theta|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\theta})] \in \mathcal{Y}$. It follows from (10) in Corollary 2.6 that the Joint-EO method yields $\hat{\boldsymbol{y}}^{J,Q}(\bar{\boldsymbol{x}}) = \mathbb{E}_{\theta|\bar{x}}[\boldsymbol{y}^*(\boldsymbol{\theta})]$. □

Appendix B. Some Conjugate Distributions.

B.1. Normal likelihood with normal prior.

- Likelihood: $x|\mu \sim \mathrm{N}(\mu,\sigma^2)$, assuming that σ^2 is known

- Prior: $\mu \sim \mathrm{N}(\mu_0, \delta^2)$
- Posterior: $\mu|\bar{x} \sim \mathrm{N}(\rho\mu_0 + (1-\rho)\bar{x}, \sigma^2(1-\rho))$ with $\rho := \frac{\sigma^2}{\sigma^2+\delta^2}$
- Posterior predictive: $x|\bar{x} \sim \mathrm{N}(\rho\mu_0 + (1-\rho)\bar{x}, \sigma^2(2-\rho))$

B.2. Exponential likelihood with gamma prior.

- Likelihood: $x|\lambda \sim \mathrm{Exp}(\lambda)$
- Prior: $\lambda \sim \mathrm{Gamma}(\alpha, \beta)$, where the order of the parameters is shape and rate, respectively
- Posterior: $\lambda|\bar{x} \sim \mathrm{Gamma}(\alpha+1, \beta+\bar{x})$
- Posterior predictive: $x|\bar{x} \sim \mathrm{Lomax}(\beta+\bar{x}, \alpha+1)$, where the order of the parameters is scale and shape, respectively

B.3. Geometric likelihood with beta prior.

- Likelihood: $x|p \sim \mathrm{Geo}(p)$
- Prior: $p \sim \mathrm{Beta}(\alpha, \beta)$
- Posterior: $p|\bar{x} \sim \mathrm{Beta}(\alpha+1, \beta+\bar{x}-1)$
- Posterior predictive: For $x = 0, 1, \ldots$, we have

$$h(x|\bar{x}) = \frac{\alpha+1}{\alpha+\beta+\bar{x}+x+2}\,\frac{\beta+\bar{x}+x}{\alpha+\beta+\bar{x}+x+1}\,\frac{\beta+\bar{x}+x-1}{\alpha+\beta+\bar{x}+x}. \tag{30}$$

REFERENCES

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

[2] A. Ben-Tal, L. Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics, 2009.

[3] A. Ben-Tal and A. Nemirovski. Robust optimization – methodology and applications. *Mathematical Programming*, 92:453–480, 2002.

[4] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53:464–501, 2011.

[5] J. Birge. Uses of sub-sample estimates in stochastic optimization models. *Operations Research*, 2016.

[6] J. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer, 2011.

[7] S. Boyd, S.-J. Kim, L. Vandenberghe, and A. Hassibi. A tutorial on geometric programming. *Optimization and Engineering*, 8:67–127, 2007.

[8] M. Charikar, C. Chekuri, and M. Pál. Sampling bounds for stochastic optimization. In *Proceedings of APPROX-RANDOM*, pages 257–269, 2005.

[9] L. Y. Chu, J. G. Shanthikumar, and Z.-J. M. Shen. Solving operational statistics via a Bayesian analysis. *Operations Research Letters*, 36:110–116, 2008.

[10] D. Davarnia and G. Cornuéjols. From estimation to optimization via shrinkage. *Operations Research Letters*, 45:642–646, 2017.

[11] M. del Mar Hershenson, S. P. Boyd, and T. H. Lee. Optimal design of a cmos op-amp via geometric programming. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 20:1–21, 2006.

[12] K. Dhamdhere, R. Ravi, and M. Singh. On two-stage stochastic minimum spanning trees. In *Proceedings of IPCO*, pages 321–334, 2005.

[13] P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *Annals of Statistics*, 7:269–281, 1979.

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

[15] C. F. Gauss. *Theoria combinationis obsercationunt erronbus minimis obnoxiae*. 1821.

[16] N. Ho-Nguyen and F. Kılınç-Karzan. Accelerating optimization under uncertainty via online convex optimization. `http://www.optimization-online.org/DB_HTML/2016/08/5571.html`, 2016.

[17] H. Jiang and U. V. Shanbhag. On the solution of stochastic optimization and variational problems in imperfect information regimes. *SIAM Journal on Optimization*, 26:2394–2429, 2016.

[18] A. J. Kleywegt, A. Shapiro, and T. Homem-De-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12:479–502, 2001.

[19] N. Kohli and J. R. Harring. Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research*, 48:370–397, 2013.

[20] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer-Verlag, 1998.

[21] R. Levi, R. Roundy, and D. Shmoys. Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research*, 32:821–839, 2007.

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

[23] A. E. B. Lim, J. G. Shanthikumar, and Z. J. M. Shen. Model uncertainty, robust optimization, and learning. In M. P. Johnson, B. Norman, and N. Secomandi, editors, *Tutorials in Operations Research: Models, Methods, and Applications for Innovative Decision Making*, pages 66–94. INFORMS, 2006.

[24] L. H. Liyanage and J. G. Shanthikumar. A practical inventory control policy using operational statistics. *Operations Research Letters*, 33:341–348, 2005.

[25] H. M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. John Wiley, 1959.

[26] A. Shapiro. Monte Carlo sampling methods. In A. Ruszczyński and A. Shapiro, editors, *Stochastic Programming, Handbooks in Operations Research and Management Science*, volume 10, pages 352–425. Elsevier, 2003.

[27] A. Shapiro and T. Homem de Mello. On the rate of convergence of Monte Carlo approximations of stochastic programs. *SIAM Journal on Optimization*, 11:70–86, 2000.

[28] A. Shapiro and A. Nemirovski. On complexity of stochastic programming problems. In V. Jeyakumar and A. Rubinov, editors, *Continuous optimization: Current Trends and Modern Applications*, volume 99, pages 111–146. Springer US, 2005.

[29] C. Swamy and D. B. Shmoys. Sampling-based approximation algorithms for multi-stage stochastic optimization. In *Proceedings of the 46th Annual IEEE Symposium on the Foundations of Computer Science*, volume 1, pages 357–366. IEEE, 2005.

[30] R. W. Thomas, D. H. Friend, L. A. Dasilva, and A. B. Mackenzie. Cognitive networks: adaptation and learning to achieve end-to-end performance objectives. *IEEE Communications Magazine*, 44:51–57, 2006.

[31] S. W. Wallace and S.-E. Fleten. Stochastic programming models in energy. *Handbooks in operations research and management science*, 10:637–677, 2003.

[32] W. T. Ziemba and R. G. Vickson. *Stochastic optimization models in finance*, volume 1. World Scientific, 2006.