Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

A STATISTICAL PERSPECTIVE ON LINEAR PROGRAMS WITH UNCERTAIN PARAMETERS

L. Jeff Hong

Department of Economics and Finance
Department of Management Sciences
City University of Hong Kong
Kowloon Tong, Hong Kong, China

Henry Lam

Department of Industrial and Operations Engineering
University of Michigan
1205 Beal Ave.
Ann Arbor, MI 48109, USA

ABSTRACT

We consider linear programs where some parameters in the objective functions are unknown but data are available. For a risk-averse modeler, the solutions of these linear programs should be picked in a way that can perform well for a range of likely scenarios inferred from the data. The conventional approach uses robust optimization. Taking the optimality gap as our loss criterion, we argue that this approach can be high-risk, in the sense that the optimality gap can be large with significant probability. We then propose two computationally tractable alternatives: The first uses bootstrap aggregation, or so-called bagging in the statistical learning literature, while the second uses Bayes estimator in the decision-theoretic framework. Both are simulation-based schemes that aim to improve the distributional behavior of the optimality gap by reducing its frequency of hitting large values.

1 INTRODUCTION

In any real applications, the input parameters of decision-making optimization models are unknown and need to be estimated from data. From a risk perspective, the solutions of such optimizations should be picked such that they can perform well over a range of scenarios as inferred from data. Our focus here, on a high level, is to find strategies to pick such solutions via a systematic use of risk criteria.

More concretely, we concentrate on deterministic linear programs (LP) in which some coefficients in the objective are unknown, but the constraints are fully known. This setup entails that the uncertainty is only on the objective function and not the feasibility of solutions. For ease of explanation, throughout most of the paper we consider the example

$$\begin{array}{ll} \max & Z(x;\theta) = \theta x_1 + x_2 \\ \text{subject to} & x_1 \le 1 \\ & x_1 + x_2 \le 2 \\ & x_1, x_2 \ge 0 \end{array} \tag{1}$$

where $x = (x_1, x_2) \in \mathbb{R}^2$ are the decision variables, and $\theta \in \mathbb{R}$ is uncertain. We assume the unknown true value of θ is $\theta_0 = 2$, and we have i.i.d. data $Y = (Y_1, \ldots, Y_n)$ generated from the distribution $N(\theta_0, \sigma^2)$ with say a known $\sigma = 5$. For convenience we also let $\mathscr{A}$ denote the fully known feasible region in (1).

To evaluate the quality of solution, we use the optimality gap as a criterion. Without fixing the parameter value at θ_0, the optimality gap of an adopted solution $\hat{x} = (\hat{x}_1, \hat{x}_2)$, as a function of θ, is given by

$$G(\hat{x}, \theta) = Z(x^*(\theta); \theta) - Z(\hat{x}; \theta)$$

where $x^*(\theta)$ denotes the optimal solution for (1) as a function of θ. Ideally, we want $G(\hat{x},\theta)$ to be small for $\theta=\theta_0$, but the true value θ_0 is never known. The main task is therefore to find procedures that are guaranteed to have small $G(\hat{x},\theta_0)$. For convenience, we denote $G_0(x)=G(x;\theta_0)$.

For (1), the true optimal solution can be easily seen to be $x^*(2)=(1,1)$ (via graphical method in Figure 1 for instance) and the optimal value is 3. Hence the optimality gap for a solution $\hat{x}$ at the true parameter value $\theta_0=2$ is $G_0(\hat{x})=3-2\hat{x}_1-\hat{x}_2$.

We stress our viewpoint of evaluating *procedures* that output some solution $\hat{x}$. A procedure will take in the input data Y, and output $\hat{x}=\hat{x}(Y)$. The distribution of $G_0(\hat{x}(Y))$ will provide insight of the quality of the procedure. The randomness in $G_0(\hat{x}(Y))$ comes from the stochasticity of Y according to the distribution $N(\theta_0,\sigma^2)$. Loosely speaking, a procedure is good if the distribution of $G_0(\hat{x}(Y))$ is concentrated at zero, whereas it is considered risky if $G_0(\hat{x}(Y))$ is large with significant probability.

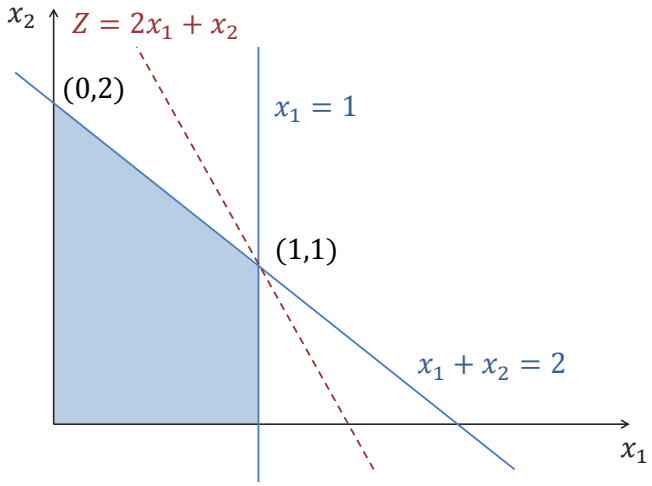

Figure 1: Graphical method for program (1).

2 CONVENTIONAL APPROACHES

2.1 Plug-in Procedure

The most basic approach for solving (1) is a "plug-in" procedure: Take the sample average of Y, namely $\bar{Y}=(1/n)\sum_{i=1}^{n}Y_i$, and use it as an estimate for θ, i.e. $\hat{\theta}=\bar{Y}$. Then solve $\max_{x\in\mathscr{X}}Z(x;\hat{\theta})$ to get $\hat{x}$.

Figure 2 shows the distribution of $G_0(\hat{x})$ using the plug-in procedure. We assume a sample size $n=20$. We generate the histogram by repeating 100 times of sampling a data set of size 20 and carrying out the plug-in scheme, and at the end finding the frequency distribution of the 100 resulting optimality gaps. As we can see, the plug-in procedure either gives perfect solution or is quite far off: Depending on the realization of the data Y, the optimality gap is either 0 or 1. This bimodal behavior is due to the solution nature of LP: an optimal solution in an LP is located at one of the corner points of the polyhedral feasible region. We classify the values of θ that lead to the same corner point optimal solution as lying in the same "decision region". It is then easy to see that there are only two decision regions for the program (1), corresponding to the solutions $(1,1)$ and $(0,2)$. The optimality gap is zero if $\hat{\theta}$ lies in the first decision region, i.e. the same one as θ_0, whereas the gap is 1 if $\hat{\theta}$ lies in the second region.

For a risk-averse modeler, this behavior of the distribution of $G_0(\hat{x})$ is arguably not satisfactory, because it implies that the loss can be substantial once a wrong solution is chosen.

2.2 Robust Optimization

Robust optimization (RO) has been widely studied in recent years (e.g. Ben-Tal et al. 2009, Bertsimas et al. 2011). Motivated by the uncertainty in the parameter θ, the idea is to find a solution $\hat{x}$ that is guaranteed to work well for a range that θ_0 is likely to lie in. This can be posted as maximizing the worst-case performance of Z as

$$\max_{x\in\mathscr{A}} \min_{\theta\in\mathscr{U}} Z(x;\theta) \tag{2}$$

where $\mathscr{U}$ denotes the *uncertainty set*. In the data-driven robust optimization framework (e.g. Bertsimas et al. 2013, Delage and Ye 2010), a common way to calibrate $\mathscr{U}$ is to use the interval estimator for θ, i.e.

$$\mathscr{U} = [\underline{\theta}, \overline{\theta}] = \left[\bar{Y} - z_\alpha \frac{\sigma}{\sqrt{n}}, \bar{Y} + z_\alpha \frac{\sigma}{\sqrt{n}}\right] \tag{3}$$

where z_α is the $1-\alpha$ quantile of the standard normal distribution. Typical value of α is 0.05. In this case the output $\hat{x}$ from (2) will guarantee that the true objective value evaluated at $\hat{x}$ is better than the optimal value of (2) with 95% confidence. From a statistical viewpoint, however, this data-driven RO approach does not correspond to any standard statistical procedure in a decision-theoretic sense, e.g. it is not a minimax estimator (Cox and Hinkley 1979; also see Section 5) for any particular loss function.

Figure 3 shows the distribution of $G_0(\hat{x})$ by using the data-driven RO formulation (2). We can see that the distribution is still bimodal; in fact, it appears even worse than the plug-in procedure as there is a higher chance that the optimality gap is 1. This shows that with respect to the optimality gap, RO is also a high-risk procedure.

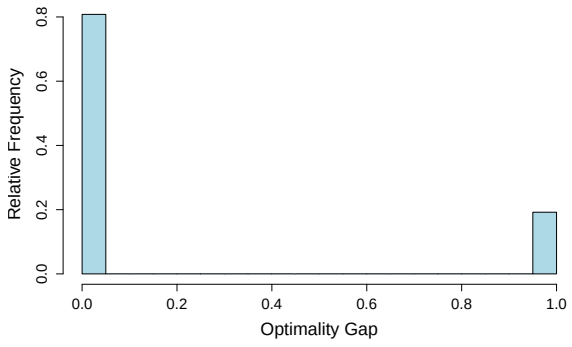

Figure 2: Histogram of optimality gap for plug-in procedure.

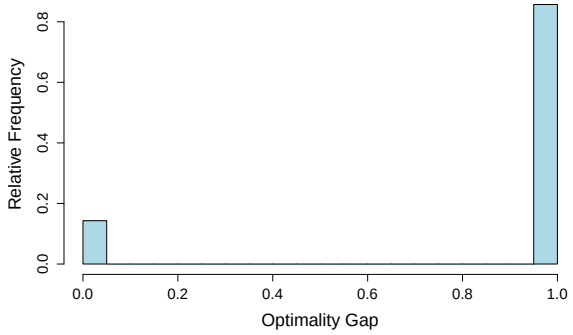

Figure 3: Histogram of optimality gap for RO.

2.3 A Comparison of Plug-in Procedure and Robust Optimization

We proceed with some mathematical analysis. Let us compute the exact distribution of $G_0(\hat{x})$ for the plug-in and RO procedure. First, we write down the function $x^*(\theta)$ as

$$x^*(\theta) = \begin{cases} (1,1) & \text{if } \theta \geq 1 \\ (0,2) & \text{if } \theta \leq 1. \end{cases} \tag{4}$$

This can be easily seen by scrutinizing Figure 1. Fixing the feasible region, the optimal solution is $(1,1)$ when the slope of the objective line is less than -1, while it is $(0,2)$ if the slope is greater than -1.

For the plug-in procedure, (4) implies that we would choose

$$\hat{x} = \begin{cases} (1,1) & \text{if } \hat{\theta} \geq 1 \\ (0,2) & \text{if } \hat{\theta} \leq 1. \end{cases}$$

Now, using the assumption that $Y_i \sim N(\theta_0, \sigma^2)$ and so $\bar{Y} \sim N(\theta_0, \sigma^2/n)$, we have

$$P(\hat{\theta} \geq 1) = P(\bar{Y} \geq 1) = \overline{\Phi}\left(\frac{1-\theta_0}{\sigma/\sqrt{n}}\right) = \overline{\Phi}\left(-\frac{\sqrt{n}}{5}\right)$$

and similarly

$$P(\hat{\theta} \leq 1) = \Phi\left(-\frac{\sqrt{n}}{5}\right)$$

where $\Phi(\cdot)$ and $\overline{\Phi}(\cdot)$ denote respectively the distribution function and the tail distribution function of standard normal variable. Therefore, we have

$$\hat{x} = \begin{cases} (1,1) & \text{with probability } \overline{\Phi}\left(-\frac{\sqrt{n}}{5}\right) \\ (0,2) & \text{with probability } \Phi\left(-\frac{\sqrt{n}}{5}\right). \end{cases}$$

Translating into optimality gap, we get

$$G_0(\hat{x}) = \begin{cases} 0 & \text{with probability } \overline{\Phi}\left(-\frac{\sqrt{n}}{5}\right) \\ 1 & \text{with probability } \Phi\left(-\frac{\sqrt{n}}{5}\right). \end{cases} \tag{5}$$

Next we turn to RO. Note that taking the uncertainty set $\mathscr{U}$ as the interval estimate in (3), we have

$$\max_{x \in \mathscr{A}} \min_{\theta \in \mathscr{U}} Z(x;\theta) = \max_{x \in \mathscr{A}} Z(x;\underline{\theta}).$$

From (4), we therefore have

$$\hat{x} = \begin{cases} (1,1) & \text{if } \underline{\theta} \geq 1 \\ (0,2) & \text{if } \underline{\theta} \leq 1. \end{cases}$$

Since

$$P(\underline{\theta} \geq 1) = P\left(\bar{Y} - z_\alpha \frac{\sigma}{\sqrt{n}} \geq 1\right) = \overline{\Phi}\left(\frac{1 + z_\alpha \sigma/\sqrt{n} - \theta_0}{\sigma/\sqrt{n}}\right) = \overline{\Phi}\left(-\frac{\sqrt{n}}{5} + z_\alpha\right)$$

and

$$P(\underline{\theta} \leq 1) = \Phi\left(-\frac{\sqrt{n}}{5} + z_\alpha\right),$$

we get

$$\hat{x} = \begin{cases} (1,1) & \text{with probability } \overline{\Phi}\left(-\frac{\sqrt{n}}{5} + z_\alpha\right) \\ (0,2) & \text{with probability } \Phi\left(-\frac{\sqrt{n}}{5} + z_\alpha\right) \end{cases}$$

and

$$G_0(\hat{x}) = \begin{cases} 0 & \text{with probability } \overline{\Phi}\left(-\frac{\sqrt{n}}{5} + z_\alpha\right) \\ 1 & \text{with probability } \Phi\left(-\frac{\sqrt{n}}{5} + z_\alpha\right). \end{cases} \tag{6}$$

Comparing (6) with (5), we see that RO has a less favorable optimality gap distribution in this setting since $\overline{\Phi}(-\sqrt{n}/5 + z_\alpha) < \overline{\Phi}(-\sqrt{n}/5)$. In fact, simple further investigation can conclude that the optimality gap from RO is at best as good as that from plug-in for any realization of Y (and strictly worse with positive probability). However, we should point out that here the true parameter value $\theta_0 = 2$ is an unlucky scenario for RO. If θ_0 had been $1/2$ for instance, then RO would have performed better than plug-in. Nevertheless, the expressions (5) and (6) reveal why the optimality gap distributions for plug-in and RO are both bimodal and less than satisfactory.

3 A MINIMAX PROCEDURE ON THE OPTIMALITY GAP

The reason why RO does not perform well with respect to optimality gap is because, by its own construction, the procedure does not take into account $G(\hat{x};\theta)$ as a risk criterion. To remedy this issue, one can consider alternately

$$\min_{x\in\mathscr{A}} \max_{\theta\in\mathscr{U}} Z(x^*(\theta);\theta) - Z(x;\theta). \tag{7}$$

The procedure (7) minimizes the worst-case optimality gap over the uncertainty set of θ. The procedure thus guarantees that the chosen $\hat{x}$ will perform at worst as the optimal value of (7), in terms of optimality gap, with 95% probability. The idea of (7) is similar to that in Lim et al. (2012) and Lim et al. (2006), which focuses on the uncertainty of parametric distributional model for operations management and finance problems. Adopting their terminology, we shall call (7) the benchmarking procedure.

To solve (7), note first that $Z(x^*(\theta);\theta)$, as a function of θ, is given by

$$Z(x^*(\theta);\theta) = \begin{cases} \theta+1 & \text{if } \theta \geq 1 \\ 2 & \text{if } \theta \leq 1 \end{cases}$$

by putting (4) into the objective function in (1). Figure 4 shows $Z(x^*(\theta);\theta)$. When the set $\mathscr{U} = [\underline{\theta}, \overline{\theta}]$ lies completely below the threshold $\theta = 1$, obviously one should choose $\hat{x} = (0,2)$, whereas when $\mathscr{U}$ lies completely above the threshold $\theta = 1$, then one should choose $\hat{x} = (1,1)$. In both cases, the worst-case optimality gap given by (7) is zero.

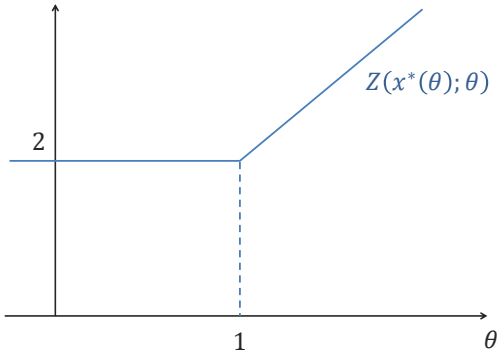

Figure 4: The function $Z(\hat{x};\theta)$ against θ.

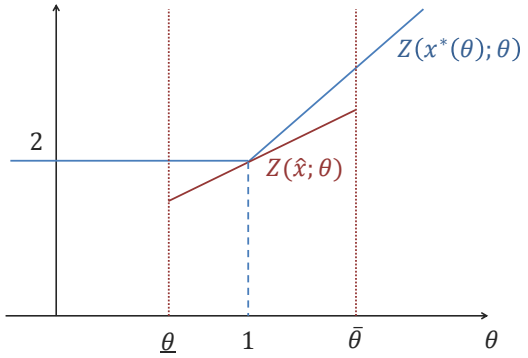

Figure 5: The function $Z(\hat{x};\theta)$ when $\mathscr{U}$ covers $\theta = 1$ in the benchmarking procedure.

The more interesting scenario is when $\mathscr{U}$ neither lies completely below or above $\theta = 1$. In this case, one can interpret (7) as trying to find a straight line, $Z(x;\theta) = \theta x_1 + x_2$ as a function of θ, that lies under the function $Z(x^*(\theta);\theta)$ such that the maximum shortfall over the range $[\underline{\theta}, \overline{\theta}]$ is minimized. Figure 5 shows such a line. We demonstrate how to find this line mathematically. First, it must touch the point $(1,2)$ since otherwise one can always lift up the line so that the maximum shortfall is decreased. This translates to the condition $x_1 + x_2 = 2$. With this constraint, the maximum shortfall must occur at the boundary of the interval $\mathscr{U}$, i.e. $\underline{\theta}$ and $\overline{\theta}$. The shortfall at $\overline{\theta}$ is $\overline{\theta} + 1 - \overline{\theta}x_1 - x_2 = (1-\overline{\theta})x_1 + (\overline{\theta} - 1)$ by substituting $x_2 = 2 - x_1$, and the shortfall at $\underline{\theta}$ is $2 - \underline{\theta}x_1 - x_2 = (1-\underline{\theta})x_1$. Therefore, by noting also that $0 \leq x_1 \leq 2$ due to the non-negativity constraint in (1), the problem here becomes solving $\min_{0\leq x_1\leq 2} \max\{(1-\overline{\theta})x_1 + (\overline{\theta}-1), (1-\underline{\theta})x_1\}$. One can draw a graph of $\max\{(1-\overline{\theta})x_1 + (\overline{\theta}-1), (1-\underline{\theta})x_1\}$ against x_1 and see easily that the minimizer occurs when $(1-\overline{\theta})x_1 + (\overline{\theta}-1)$ intersects $(1-\underline{\theta})x_1$. Thus, setting $(1-\overline{\theta})x_1 + (\overline{\theta}-1) = (1-\underline{\theta})x_1$, we have

$$x_1 = \frac{\overline{\theta} - 1}{\overline{\theta} - \underline{\theta}}$$

which is always between 0 and 1 if $\overline{\theta} \geq 1$ and $\underline{\theta} \leq 1$ (and $\overline{\theta} \neq \underline{\theta}$), or in other words $\mathscr{U}$ does not lie completely below or above the threshold $\theta = 1$.

Therefore, we have

$$\hat{x} = \begin{cases} (1,1) & \text{if } \mathscr{U} \subset [1,\infty) \\ (0,2) & \text{if } \mathscr{U} \subset (-\infty,1] \\ \left(\frac{\overline{\theta}-1}{\overline{\theta}-\underline{\theta}}, 2-\frac{\overline{\theta}-1}{\overline{\theta}-\underline{\theta}}\right) & \text{otherwise.} \end{cases} \tag{8}$$

Figure 6 shows the distribution of the optimality gap for the benchmarking procedure (7). We can see that the distribution now is more spread out between the two extremes 0 and 1, and has a substantially lower probability of having a large optimality gap. In fact, the probability of having an optimality gap at around 1 drops from 0.2 in the case of plug-in and 0.8 in the case of RO to close to 0.01 for benchmarking. The change in the shape of the optimality gap distribution comes from the phenomenon that the solution (8) is no longer concentrated at the corner points $(1,1)$ and $(0,2)$. Rather, there is a smooth transition from $(1,1)$ to $(0,2)$ as the uncertainty set moves from $[1,\infty)$ to $(-\infty,1]$ (as a comparison, Figure 7 shows the only two possibilities of $Z(\hat{x};\theta)$ in the plug-in or RO procedure). In general, this smoothing of the decision at the boundary between decision regions appears to help flatten the optimality gap distribution from being concentrated at extremes.

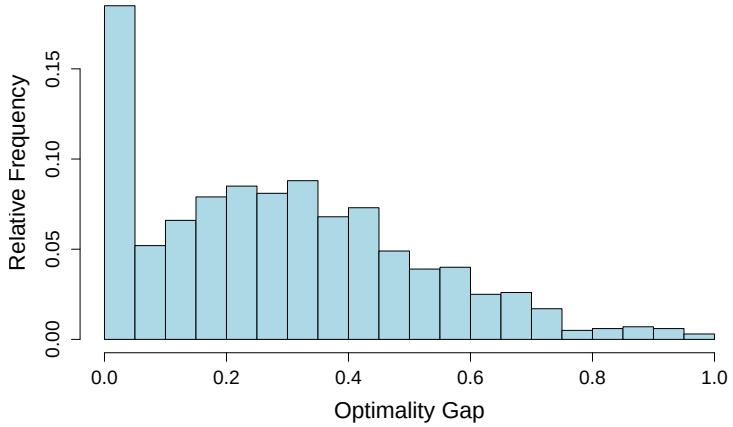

Figure 6: Histogram of optimality gap for the benchmarking procedure.

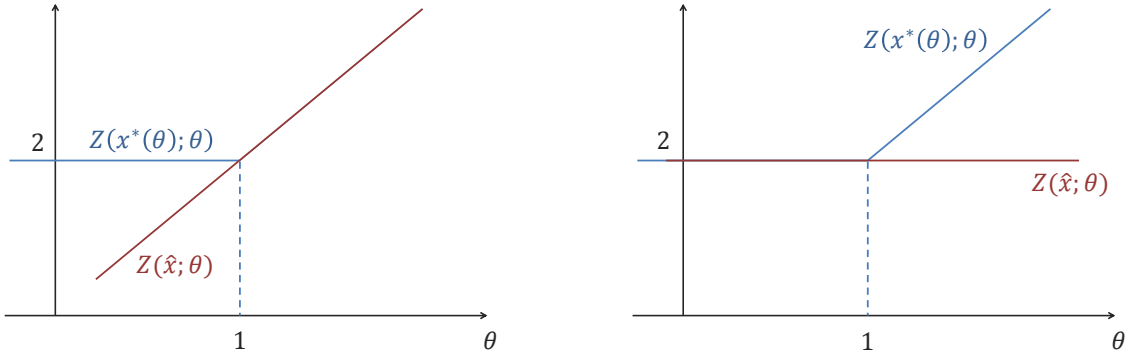

Figure 7: Two only possibilities of $Z(\hat{x};\theta)$ under the plug-in or RO procedure.

Unfortunately, (7) is not a convex program, and so it raises the question of whether it can be solved for more general problems. In the next sections, we will look at two alternate procedures that have a similar effect as (7) but computationally more tractable.

4 BOOTSTRAP AGGREGATION

4.1 Procedure and Empirical Performance

Bootstrap aggregation, or what is known as bagging, is a technique originated from classification problems that is used to reduce variance and avoid overfitting (Breiman 1996). It is now a widely used technique embedded in some off-the-shelf machine learning algorithms like random forest (Hastie et al. 2009), and is known to improve estimation accuracy in settings of joint statistical estimation and model selection (Efron 2014). The main idea of bagging is to resample the data set and repeat the estimation procedures, and at the end take a sample average of the resampled estimators.

In our setting, bagging consists of:

1. Generate n samples with replacement from $\{Y_1,\ldots,Y_n\}$. Call them $Y_1^b,\ldots,Y_n^b$. Then solve the plug-in optimization $\max_{x\in\mathscr{A}} Z(x;\bar{Y}^b)$ where $\bar{Y}^b$ is the sample average of Y_i^b's.
2. Repeat the above B times. Let the B optimal solutions be $x^1,x^2,\ldots,x^B$.
3. Output $\hat{x} = (1/B)\sum_{j=1}^{B} x^j$.

Figure 8 shows the distribution of $G_0(\hat{x})$ by applying bagging with $B = 100$. The distribution looks similar to that in Figure 6, i.e. benchmarking, in that they are both spread out in between the extremes 0 and 1. The probability for the optimality gap being close to 1 in Figure 8 is kept at a low number of 0.02. In fact, its shape looks even better than that in Figure 6, in the sense that higher frequency occurs at small optimality gaps. Computationally, bagging requires solving B number of LPs with the same complexity.

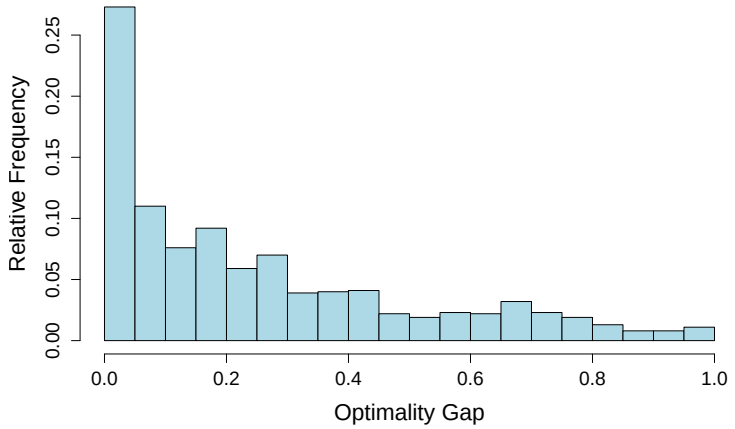

Figure 8: Histogram of optimality gap for the bagging procedure.

4.2 Mathematical Explanation

We discuss how bagging works from a statistical perspective. We decompose the mean square of the optimality gap, $E[G_0(\hat{x})^2]$, into a bias term and a variance term, given by $E[G_0(\hat{x})^2] = (E[G_0(\hat{x})])^2 + Var(G_0(\hat{x}))$. Compared to the plug-in procedure, what bagging does is to reduce the variance while retaining a similar bias, and hence the overall value of $E[G_0(\hat{x})^2]$ is lower than that of the plug-in procedure. This argument

borrows from the theoretical study of bagging in the setting of classification. In the latter scenario, the loss function is typically discrete, e.g. 0 or 1 corresponding to correct or incorrect classification, and bagging improves variance by "smoothing" the loss function at the transition boundary between decision regions. Translating to the LP setting, one can think of LP as a classification problem where the classification outcomes are exactly the corner points that are possibly optimal.

To illustrate the above discussion in more detail, we shall show some heuristic calculation to compare bagging with the plug-in procedure. The argument we shall use borrows largely from the technique in Büchlmann and Yu (2002). For the rest of this section only, we focus on a simpler LP with one decision variable:

$$\begin{array}{ll} \max & \theta x \\ \text{subject to} & a \le x \le b. \end{array} \tag{9}$$

Say $\theta_0 > 0$. The true optimal solution is hence b. Again, we assume that i.i.d. data $Y = (Y_1, \ldots, Y_n)$ are available, with $Y_i \sim N(\theta_0, \sigma^2)$. For the plug-in procedure, we have $\hat{x} = bI(\hat{\theta} \ge 0) + aI(\hat{\theta} < 0)$, where $I(\cdot)$ is the indicator function. Hence the bias is

$$\theta_0 E[b - (bI(\hat{\theta} \ge 0) + aI(\hat{\theta} < 0))] = \theta_0 (b-a) P(\hat{\theta} < 0).$$

Suppose $\theta_0 = c/n^\alpha$ for some constant $c > 0$, where n is the number of data. We distinguish three cases: $0 \le \alpha < 1/2$, $\alpha = 1/2$ and $\alpha > 1/2$.

Note that $\sqrt{n}(\hat{\theta} - \theta_0)/\sigma \overset{\text{approx.}}{\sim} N(0,1)$. Writing $P(\hat{\theta} < 0) \approx \Phi(-\sqrt{n}\theta_0/\sigma)$, we have:

1. If $0 \le \alpha < 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to \infty \;\Rightarrow P(\hat{\theta} < 0) \approx \frac{\sigma}{\sqrt{2\pi n}\theta_0} e^{-n\theta_0^2/(2\sigma^2)}$
   $\Rightarrow$ bias has exponential decay rate $\theta_0^2/(2\sigma^2)$ as $n \to \infty$.
2. If $\alpha = 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to c \;\Rightarrow P(\hat{\theta} < 0) \to \overline{\Phi}(c) \;\Rightarrow$ bias $\to \theta_0(b-a)\overline{\Phi}(c)$.
3. If $\alpha > 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to 0 \;\Rightarrow P(\hat{\theta} < 0) \to \frac{1}{2} \;\Rightarrow$ bias $\to \theta_0(b-a)\frac{1}{2}$.

Similarly, we can derive that the variance for plug-in is ${\theta_0}^2(b-a)^2 P(\hat{\theta} < 0)P(\hat{\theta} \ge 0)$, and we have:

1. If $0 \le \alpha < 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to \infty \;\Rightarrow$ variance has exponential decay rate ${\theta_0}^2/(2\sigma^2)$ as $n \to \infty$.
2. If $\alpha = 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to c \;\Rightarrow$ variance $\to {\theta_0}^2(b-a)^2\overline{\Phi}(c)\Phi(c)$.
3. If $\alpha > 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to 0 \;\Rightarrow$ variance $\to {\theta_0}^2(b-a)^2\frac{1}{4}$.

Now consider the case of bagging. For convenience we assume we can take the number of resamples $B = \infty$. The adopted solution in this case is $\hat{x} = \tilde{E}[bI(\tilde{\theta} \ge 0) + aI(\tilde{\theta} < 0)] = b\tilde{P}(\tilde{\theta} \ge 0) + a\tilde{P}(\tilde{\theta} < 0)$, where $\tilde{\theta}$ denotes a bootstrapped estimate of θ given Y and $\tilde{E}$ and $\tilde{P}$ denote respectively the expectation and probability with respect to $\tilde{\theta}$, which is distributed approximately as $N(\hat{\theta}, \sigma^2/n)$ with $\hat{\theta} = \bar{Y}$. Consider the optimality gap

$$\begin{aligned} \theta_0(b - b\tilde{P}(\tilde{\theta} \ge 0) - a\tilde{P}(\tilde{\theta} < 0)) &= \theta_0(b-a)\tilde{P}(\tilde{\theta} < 0) \\ &= \theta_0(b-a)\overline{\Phi}\left(\frac{\sqrt{n}\hat{\theta}}{\sigma}\right) \\ &= \theta_0(b-a)\overline{\Phi}\left(\frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sigma} + \frac{\sqrt{n}\theta_0}{\sigma}\right). \end{aligned}$$

The bias is given by

$$\theta_0(b-a)E\left[\overline{\Phi}\left(\frac{\sqrt{n}(\hat{\theta}-\theta_0)}{\sigma}+\frac{\sqrt{n}\theta_0}{\sigma}\right)\right]$$

where the expectation E is on $\hat{\theta}$. Since $\frac{\sqrt{n}(\hat{\theta}-\theta_0)}{\sigma} \sim N(0,1)$, we have:

1. If $0 \leq \alpha < 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to \infty \Rightarrow$ bias has exponential decay rate ${\theta_0}^2/(2\sigma^2)$ as $n \to \infty$.
2. If $\alpha = 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to c \Rightarrow$ bias $\to \theta_0(b-a)E[\overline{\Phi}(Z+c)]$ where $Z \sim N(0,1)$.
3. If $\alpha > 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to 0 \Rightarrow$ bias $\to \theta_0(b-a)E[\overline{\Phi}(Z)] = \theta_0(b-a)E[U] = \theta_0(b-a)\frac{1}{2}$ where $U \sim Unif(0,1)$.

Similarly, the variance is

$${\theta_0}^2(b-a)^2 Var\left(\overline{\Phi}\left(\frac{\sqrt{n}(\hat{\theta}-\theta_0)}{\sigma}+\frac{\sqrt{n}\theta_0}{\sigma}\right)\right).$$

Hence:

1. If $0 \leq \alpha < 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to \infty \Rightarrow$ variance has exponential decay rate ${\theta_0}^2/\sigma^2$ as $n \to \infty$.
2. If $\alpha = 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to c \Rightarrow$ variance $\to {\theta_0}^2(b-a)^2 Var(\overline{\Phi}(c+Z))$.
3. If $\alpha > 1/2$, then $\frac{\sqrt{n}\theta_0}{\sigma} \to 0 \Rightarrow$ variance $\to {\theta_0}^2(b-a)^2 Var(\overline{\Phi}(Z)) = {\theta_0}^2(b-a)^2 Var(U) = {\theta_0}^2(b-a)^2\frac{1}{12}$.

We argue that for all three cases $0 \leq \alpha < 1/2$, $\alpha = 1/2$ and $\alpha > 1/2$, bagging provides advantages or is roughly as good as plug-in. If θ_0 is close to 0 relative to the sample size, i.e. $\alpha > 1/2$, then the limiting variance is strictly smaller in bagging than in plug-in since ${\theta_0}^2(b-a)^2\frac{1}{12} < {\theta_0}^2(b-a)^2\frac{1}{4}$. On the other hand, the limiting biases are the same at $\theta_0(b-a)\frac{1}{2}$ for both bagging and plug-in. For the case $\alpha < 1/2$, both bias and variance have exponential decay in n in both bagging and plug-in, and the difference between the two schemes is thus negligible. In the case $\alpha = 1/2$, bagging often leads to a reduction in variance while maintaining similar bias as plug-in. An illustrative scenario is when $c = 0$, which reduces to the same conclusion as the case for $\alpha > 1/2$. When $c > 0$, one can still see that bagging has a smaller overall mean square error for a large range of c (Figure 2 in Büchlmann and Yu 2002). We note that Büchlmann and Yu (2002) focuses on the case $\alpha > 1/2$ in their analysis, and we have considered the additional cases of $0 \leq \alpha < 1/2$ and $\alpha = 1/2$ here.

5 DECISION-THEORETIC APPROACH

The next approach we consider is inspired from decision theory (e.g. Cox and Hinkley 1979). Viewing $\hat{x}$ as an estimate of the true optimal solution $x^*(\theta_0)$, we can evaluate the quality of a statistical estimation procedure by the use of risk function

$$E_{Y|\theta}[l(G(\hat{x}(Y),\theta))] \tag{10}$$

where $E_{Y|\theta}[\cdot]$ denotes the expectation taken on the data Y given a true parameter θ, and $l(\cdot): \mathbb{R}_+ \to \mathbb{R}$ is a convex non-decreasing loss function. One way to minimize (10) while taking into account the uncertainty of θ is use a minimax estimator, namely by finding $\hat{x}(Y)$ that solves

$$\min_{\hat{x}(Y)} \max_{\theta} E_{Y|\theta}[l(G(\hat{x}(Y),\theta))]. \tag{11}$$

The formulation (11) can be difficult to solve in general, because the decision variable is a procedure, not a numerical object. We should also point out that the framework in (11) is very different from data-driven RO discussed in Section 2.2, because finding a solution that minimizes the worst-case loss function over a calibrated uncertainty set for θ is not equivalent to finding a procedure that minimizes the worst-case expectation of the loss function with respect to the likelihood of the data Y.

One tractable machinery to handle (10) is to use the Bayesian framework. Consider the Bayes risk

$$E_{Y,\theta}[l(G(\hat{x}(Y),\theta))] = E_{\theta}[E_{Y|\theta}[l(G(\hat{x}(Y),\theta))]] \tag{12}$$

where $E_{\theta}[\cdot]$ is the expectation taken with respect to some prior distribution on θ. Note that (12) can be written as

$$E_{Y}[E_{\theta|Y}[l(G(\hat{x}(Y),\theta))]]$$

where now $E_{\theta|Y}[\cdot]$ denotes the posterior expectation of the parameter θ given the data Y, and $E_{Y}[\cdot]$ is taken with respect to the unconditional distribution of Y. A Bayes procedure minimizes the Bayes risk by solving

$$\min_{x\in\mathscr{A}} E_{\theta|Y}[l(G(x,\theta))]. \tag{13}$$

Supposing that the distribution of θ given Y can be computed, then (13) is a convex program. In fact, a more general theorem is available:

Theorem 1 Consider the convex optimization $\max_{x\in\mathscr{A}} f(x;\Theta)$, where f is a concave objective function in x with parameter Θ and $\mathscr{A}$ is some known convex deterministic set. The Bayes procedure $\min_{x\in\mathscr{A}} E_{\Theta|Y}[l(f(x^*(\Theta);\Theta) - f(x;\Theta))]$, where $l:\mathbb{R}_+\to\mathbb{R}$ is a convex non-decreasing loss function and $x^*(\Theta)$ is an optimal solution for parameter Θ, is also a convex program.

Proof. Since $f(x;\Theta)$ is concave, $f(x^*(\Theta);\Theta) - f(x;\Theta)$ is convex, and moreover it is non-negative by the definition of $x^*(\Theta)$. Therefore, since l is convex non-decreasing on $\mathbb{R}_+$, the quantity $E_{\Theta|Y}[l(f(x^*(\Theta);\Theta) - f(x;\Theta))]$ is convex in x. □

Moreover, for quadratic loss function, the Bayes procedure is a convex quadratic program (QP) if the original optimization is an LP:

Theorem 2 Using the notation in Theorem 1, if $l(w) = w^2$ and $\max_{x\in\mathscr{A}}\{f(x;\Theta) = \Theta' x\}$ is a linear program, then the Bayes procedure is a convex QP.

Proof. We can write

$$\begin{aligned} E_{\Theta|Y}[l(f(x^*(\Theta);\Theta) - f(x;\Theta))] &= E_{\Theta|Y}(\Theta'(x^*(\Theta) - x))^2 \\ &= E_{\Theta|Y}[(x^*(\Theta) - x)'\Theta\Theta'(x^*(\Theta) - x)] \\ &= x' E_{\Theta|Y}[\Theta\Theta']x - 2E_{\Theta|Y}[x^*(\Theta)'\Theta\Theta']x + E_{\Theta|Y}[x^*(\Theta)'\Theta\Theta' x^*(\Theta)] \end{aligned} \tag{14}$$

which is a convex quadratic form in x.

□

We carry out the Bayes procedure using a quadratic loss function for (1). Here we denote $\Theta = (\theta, 1)'$. By the proof of Theorem 2, we need to solve

$$\min_{x\in\mathscr{A}} x' E_{\theta|Y}[\Theta\Theta']x - 2E_{\theta|Y}[x^*(\Theta)'\Theta\Theta']x + E_{\theta|Y}[x^*(\Theta)'\Theta\Theta' x^*(\Theta)]. \tag{15}$$

We consider two alternatives:

1. Put a normal prior on θ, i.e. $\theta \sim N(\mu, \gamma^2)$. Then the posterior distribution is (Gelman et al. 2014)

$$\theta|Y \sim N\left(\frac{\mu/\gamma^2 + (n/\sigma^2)\bar{Y}}{1/\gamma^2 + n/\sigma^2}, \left(\frac{1}{\gamma^2} + \frac{n}{\sigma^2}\right)^{-1}\right).$$

2. Use the empirical distribution of Y to replace the posterior distribution. The empirical distribution is not a posterior distribution in theory, but is a good approximation when n is large.

Note that only the first two terms in the objective function in (15) are relevant for finding the optimal solution. The second term involves the quantity $E_{\theta|Y}[x^*(\Theta)'\Theta\Theta']$, which is in general difficult to evaluate in closed form. Therefore we approximate it by drawing samples from the posterior or the empirical distribution (depending on which alternatives above). Figures 9 and 10 show the distributions of optimality gap for using the normal prior and the empirical distribution approximation respectively. We can see that in both cases the distribution is spread out between 0 and 1, thus remedying the issue in plug-in and RO. Compared to bagging, i.e. Figure 8, the probabilities of optimality gap close to 1 are slightly higher in the two Bayes procedures, both being around 0.05, than for bagging, which is around 0.02. However, higher frequency occurs at smaller values of the optimality gap in the Bayes procedures, being between 0.4 and 0.5 for optimality gap close to zero, compared to around 0.25 for bagging.

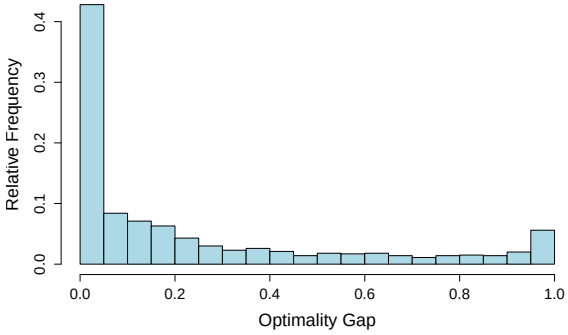

Figure 9: Histogram of optimality gap for Bayes estimator using normal prior.

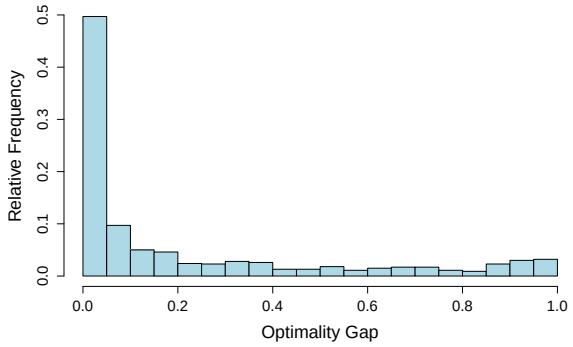

Figure 10: Histogram of optimality gap for Bayes estimator using empirical distribution approximation.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CMMI-1400391 and CMMI-1436247, and Hong Kong Research Grants Council under grant GRF9042157.

REFERENCES

Ben-Tal, A., L. El Ghaoui, and A. Nemirovski. 2009. *Robust optimization*. Princeton University Press.

Bertsimas, D., D. B. Brown, and C. Caramanis. 2011. "Theory and applications of robust optimization". *SIAM review* 53 (3): 464–501.

Bertsimas, D., V. Gupta, and N. Kallus. 2013. "Data-driven robust optimization". *arXiv preprint arXiv:1401.0212*.

Breiman, L. 1996. "Bagging predictors". *Machine learning* 24 (2): 123–140.

Büchlmann, P., and B. Yu. 2002. "Analyzing bagging". *Annals of Statistics*:927–961.

Cox, D. R., and D. V. Hinkley. 1979. *Theoretical statistics*. CRC Press.

Delage, E., and Y. Ye. 2010. "Distributionally robust optimization under moment uncertainty with application to data-driven problems". *Operations research* 58 (3): 595–612.

Efron, B. 2014. "Estimation and accuracy after model selection". *Journal of the American Statistical Association* 109 (507): 991–1007.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2014. *Bayesian data analysis*, Volume 2. Taylor & Francis.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning*, Volume 2. Springer.

Lim, A. E., J. G. Shanthikumar, and Z. M. Shen. 2006. "Model uncertainty, robust optimization, and learning". *Tutorials in Operations Research: Models, Methods, and Applications for Innovative Decision Making*:66–94.

Lim, A. E., J. G. Shanthikumar, and G.-Y. Vahn. 2012. "Robust portfolio choice with learning in the framework of regret: Single-period case". *Management Science* 58 (9): 1732–1746.

AUTHOR BIOGRAPHY

L. JEFF HONG is a chair professor in the Department of Economics and Finance, and the Department of Management Sciences, at the City University of Hong Kong. His research interests include Monte Carlo method, financial engineering, and stochastic optimization. He is currently an associate editor for *Operations Research*, *Naval Research Logistics* and *ACM Transactions and Modeling and Computer Simulation*. His email address is jeffhong@cityu.edu.hk.

HENRY LAM is an Assistant Professor in the Department of Industrial and Operations Engineering at the University of Michigan, Ann Arbor. He graduated from Harvard University with a Ph.D. degree in statistics in 2011, and has been an Assistant Professor in the Department of Mathematics and Statistics at Boston University until 2014. His research focuses on stochastic simulation, risk analysis, and simulation optimization. His email address is khlam@umich.edu.