The Value of Side Information in Shortest Path Optimization

Michael Rinehart and Munther A. Dahleh

***Abstract*—Consider an agent who seeks to traverse the shortest path in a graph having random edge weights. If the agent has no side information about the realizations of the edge weights, it should simply take the path of least average length. We consider a generalization of this framework whereby the agent has access to a limited amount of side information about the edge weights ahead of choosing a path. We define a measure for information quantity, provide bounds on the agent's performance relative to the amount of side information it receives, and present algorithms for optimizing side information. The results are based on a new graph characterization tied to shortest path optimization.**

***Index Terms*—Estimation, optimal control, optimization.**

I. Introduction

CONSIDER an agent who seeks to traverse the shortest path in a graph having random edge weights. If the agent has no information about the realizations of the edge weights, it should simply take the path of least average length (a simple optimization problem). In this paper, we consider a generalization of this framework whereby the agent can use side information about the random edge weights of a graph to determine the shortest-average path in the graph. Specifically, the agent can use side information about the edge weights to estimate the length of each path and can further optimize the side information subject to a bound on information quantity.

In our setting, the value of side information is measured by the average length of the paths the agent chooses, not how often the agent decodes the optimal path. We define a measure for information that is compatible with this problem, bound the performance of the agent subject to a bound on information quantity, and present algorithms for information optimization. Meaningful, analytic performance bounds and practical algorithms for information optimization are based in a new type of graph characterization tied to shortest path optimization.

Our formulation is a special case of stochastic optimization, and numerous papers have considered the problem of bounding the improvement one gets from information in stochastic programming. For instance, [1] studies the value of having full information by subdividing the domain computing bounds over each subdomain. [2] takes a different approach by leveraging a concavity assumption to derive a computational bound for performance, but the bound is worse than that obtained using Jensen's Inequality. In [3] and [4], the impact of partial information is considered. In the first, partial information is represented by a signal that offers information about the underlying distribution of the uncertainty, and it is assumed that there are a finite number of such distributions. [4] similarly represents partial information, but the authors seek to determine the worst case performance of the optimization over the unknown distributions.

Because the results from these works are not specific to any particular application, the bounds tend to be overly conservative and non-analytic. Because we are considering a specific formulation (shortest path optimization), we seek to provide bounds tied to the underlying structure of the problem. The performance of shortest/longest path algorithms on random graphs has been considered in the literature, but in different contexts than we seek. For instance, [5] computes the probability density function of the shortest path length in a complete graph having integer edge weights, and [6] studies the average length of the shortest path in a complete graph with uniform edge weights. Both works leverage the significant symmetry of their respective formulations, so they do not generalize to our framework. References [7] and [8], on the other hand, compute bounds on the length of the shortest path on an arbitrary graphs having arbitrary edge-weight distributions, but the bounds are independent of the graph's topology. In fact, the bound in [8] is equal to a relaxation of our bound. [9] also considers arbitrary graphs having arbitrary distributions, and achieves a topologically dependent bound by employing dynamic programming and bounding the lengths of subpaths along the dynamic program. However, the bound must be computed via an iterative algorithm.

Although not presented in the context of valuing information, the results in [10] do indirectly address portions of our framework. Specifically, the authors apply mathematical programming techniques to compute lower bounds for stochastic 1-0 optimizations. The results are based in the generalized Chebyshev bound [11], which is hard to compute in general, but [10] is able to reduce the complexity of the computation via a series of relaxations. However, these relaxations are too conservative for optimizing information. Furthermore, the computational nature of the technique does not offer insight into how topology or information quantity impact performance.

The outline of this paper is as follows. In Section II, we define the concepts and notation used in this paper, formally describe the framework for shortest path optimization under lim-

Manuscript received November 05, 2009; revised July 07, 2010 and October 29, 2010; accepted December 03, 2010. Date of publication January 13, 2011; date of current version September 08, 2011. Recommended by Associate Editor H. Ishii.

The authors are with the Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: dahleh@mit.edu; mdrine@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2011.2105744

ited information, and present the objectives of this paper. In Section III, we present a graph characterization for shortest path optimization and use it to derive information optimization algorithms and performance bounds in Sections IV and V. In Section VI, we apply our results in several examples. Finally, Section VII, we summarize the results of the paper and remark on some possible future extensions.

II. Definitions, Notation, and Formulation

A. Random Variables, Sets, and Matrices

We write random variables (RVs) in capital letters (e.g., X), and we write $X \sim p$ if X has p as its probability density function (pdf). We specifically let $N(\mu, \sigma^2)$ be the normal distribution with mean μ and variance σ^2. $\mathrm{E}[X]$ and $\mathrm{VAR}[X]$ are, respectively, the expected value and variance of X, and for a random vector $X = (X_1, \ldots, X_n)$, $\mathrm{VAR}[X] = \sum_{i=1}^{n} \mathrm{VAR}[X_i]$ whereas $\mathrm{COV}[X] = \mathrm{E}[XX^T] - \mathrm{E}[X]\mathrm{E}[X]^T$. For two RVs X, Y, we define $\hat{X}(Y) = \mathrm{E}[X|Y]$ as the estimate of X given Y (which we simplify to $\hat{X}$ if the argument is understood), and we say $X \stackrel{\mathrm{d}}{=} Y$ if both RVs are drawn from the same distribution.

If A is a set, $|A|$ is the number of elements in A. For another set B, $A \setminus B$ is the set of elements in A but not in B. If $A \subset \Re^n$ and $x \in \Re^n$, $A - x = \{a - x | a \in A\}$. We denote the sphere of radius r and center c as $B(r, c) = \{x | \; \|x - c\|^2 \leq r^2\}$.

Finally, for a positive semi-definite (PSD) matrix M, $N = \sqrt{M}$ is the unique positive (semi)definite matrix satisfying $M = N^2$, and for two PSD matrices M and N, the inequality $M \preceq N$ means that $N - M$ is PSD.

B. Graphs

We define a graph G by a pair (V, E) of vertices V and edges E. Because we allow any two vertices to have multiple edges connect them, we forgo the usual definition $E \subset V \times V$ and instead define a head and tail for each edge $e \in E$ by $\mathrm{hd}(e) \in V$ and $\mathrm{tl}(e) \in V$, respectively.

Each edge e in the graph is associated with an *edge weight* w_e. The vector of all weights is $w = [w_1 \ldots w_{|E|}]^T$. Because we consider edge weights to be random, we write the vector as W, and we assume that the probability distribution is known. Finally, let $\mu = \mathrm{E}[W]$, $\mu_e = \mathrm{E}[W_e]$, $\Lambda_W = \mathrm{COV}[W]$, and $\sigma_e^2 = \mathrm{VAR}[W_e]$.

We now define the notion of a *path* in the graph.

Definition 1 (Path): A sequence $p = (e_1, e_2, \ldots, e_n)$ with $e_i \in E$ is a path if $\mathrm{tl}(e_i) = \mathrm{hd}(e_{i+1})$, and we say p goes from $v_1 = \mathrm{hd} e_1$ to $v_{n+1} = \mathrm{tl} e_n$.

Definition 2 (Acyclic Path): A path $p = (e_i)_i$ is acyclic if there are no two indices $i < j$ such that $\mathrm{hd}(e_i) = \mathrm{tl}(e_j)$.

Assumption 1: All paths of G are acyclic (it is a directed acyclic graph).

We also assume the existence of two vertices $s, t \in V$, respectively termed the *start* and *termination* vertices, that satisfy the following assumption.

Assumption 2: There is a path from vertex s to each vertex $v \in V \setminus \{s\}$ as well as a path from each vertex $v \in V \setminus \{t\}$ to vertex t.

Let $P = P(G)$ be the set of all paths from s to t in G. From here on, when we refer to a path, it is assumed to come from P.

With some abuse of notation, we can write each $p \in P$ as a 0–1 vector in $\Re^{|E|}$, where $p_e = 1$ if $e \in p$ and $p_e = 0$ otherwise. In this case, P is also the set of all such vectors in $\Re^{|E|}$. Let $\mathcal{P} = \mathrm{convex\ hull}\{P\}$. An well-known, efficient representation for P is

$$P = x \in \{0,1\}^{|E|} \text{ such that}$$
$$\sum_{e|\mathrm{hd}(e)=v} x_e - \sum_{e|\mathrm{tl}(e)=v} x_e = \begin{cases} 1, & v = s \\ -1, & v = t \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

Finally, using our vector notation, the length of a path $p \in P$ is simply $p^{\mathrm{T}} W$.

C. Partial Information in Stochastic Optimization

We begin by developing a general formulation for studying the value of information in stochastic optimization. For the purposes of this section, let W be any RV with some known distribution.

Consider the following stochastic optimization:

$$J(W) = \min_{x \in \mathcal{X}} h(x, W).$$

Clearly, since W is a RV, $J(W)$ is also a RV, and so the average performance of the optimization is $\mathrm{E}[J(W)]$.

Consider now the task of finding an "optimal" decision x without having the realization of W. A reasonable objective is to select the x that minimizes the average of the objective:

$$\overline{J} = \min_{x \in \mathcal{X}} \mathrm{E}\left[h(x, W)\right].$$

Since $\overline{J}$ is a constant, $\mathrm{E}[\overline{J}] = \overline{J}$. By Jensen's Inequality, $\mathrm{E}[J(W)] \leq \overline{J}$.

We call the first case (where the realization of W was known) the *full-information* case. We call the latter case the *zero-information* case.

We are interested in formulating an in-between *partial-information* case. To this end, we introduce another RV Y that represents the agent's *side information* about W and write the optimization as a function of our side information:

$$J(Y) = \min_{x \in \mathcal{X}} \mathrm{E}\left[h(x, W)|Y\right].$$

Y could the output of a network of sensors on the graph's edges, for instance. Once again, because Y is a RV, $J(Y)$ is also a RV, and so the average performance under Y is simply $\mathrm{E}[J(Y)]$. Clearly, the information Y contains about W is completely determined by their joint-distribution p_{WY}, so we define $J(p_{WY}) = \mathrm{E}[J(Y)]$:

$$J(p_{WY}) = \mathrm{E}\left[\min_{x \in \mathcal{X}} \mathrm{E}\left[h(x, W)|Y\right]\right]. \quad (2)$$

Remark 1: Intuitively, we are "averaging-out" the information about $h(x, W)$ that we do not have from Y, much like in the zero-information case. The full- and zero-information cases are easily obtained by substituting $Y = 0$ (a constant) and $Y = W$ to, respectively, yield $J(Y) = \overline{J}$ and $J(Y) = J(W)$.

Proposition 1: $\mathrm{E}[J(W)] \leq J(p_{WY}) \leq \overline{J}$

Proof: The proof is a simple application of Jensen's Inequality:

$$\begin{aligned}\mathrm{E}\left[J(W)\right] &= \mathrm{E}\left[\min_{x\in X} h(x,W)\right] = \mathrm{E}\left[\mathrm{E}\left[\min_{x\in X} h(x,W)|Y\right]\right]\\ &\leq \mathrm{E}\left[\min_{x\in X} \mathrm{E}\left[h(x,W)|Y\right]\right] = J(p_{WY})\\ &\leq \min_{x\in X} \mathrm{E}\left[\mathrm{E}\left[h(x,W)|Y\right]\right] = \min_{x\in X} \mathrm{E}\left[h(x,W)\right] = \overline{J}.\end{aligned}$$

■

D. Quantifying Information and Information Optimization

The agent is also given some flexibility in determining the side information it receives in the form of being able to choose the joint distribution p_{WY}. Without added constraints, though, the agent will choose a distribution that yields $\hat{W}(Y) = W$. Therefore, we define a "bound" Γ to limit the set of allowable distributions. In terms of performance, we seek the solution to

$$J(\Gamma) = \min_{p_{WY}\in\Gamma} J(p_{WY}). \tag{3}$$

We call (3) the *information optimization*.

To quantify information, we further generalize the concept of an information bound to a family of constraint sets $\{\Gamma(C)\}$ parameterized by a non-negative scalar C called the *capacity*. For ease, we simplify our notation by writing

$$J(C) = J\left(\Gamma(C)\right)$$

and call $J(C)$ the *optimal performance under capacity* C. Although not critical to the analysis of this paper, desirable properties of $\Gamma(C)$ include the following:

- $\Gamma(C_1) \subset \Gamma(C_2)$ if $C_1 \leq C_2$;
- $p_{WY} \in \Gamma(0)$ implies $\hat{W}(Y) = \mathrm{E}[W|Y] = \mathrm{E}[W]$ for $(W,Y) \sim p_{WY}$;
- if $p_{W(Y_1 Y_2)} \in \Gamma(C)$, then p_{WY_1} and p_{WY_2} are in $\Gamma(C)$;[1]
- and there exists a C_{max} such that there exist a $p_{WY} \in \Gamma(C_{max})$ satisfying $\hat{W}(Y) = W$.

E. Partial Information in Shortest Path Optimization

We now specialize our framework to shortest path optimization. We begin by defining the information constraint sets $\{\Gamma(C)\}$:

$$\Gamma(C) = \left\{p_{WY}|\mathrm{VAR}\left[\mathrm{E}[W|Y]\right] \leq C\right\}.$$

Our choice of $\Gamma(C)$ is a practical selection motivated by the analysis that is to follow in this paper. It furthermore obeys our desired properties for $\{\Gamma(C)\}$.

Proposition 2 (Projection Theorem): $0 = \mathrm{VAR}[\mathrm{E}[W]] \leq \mathrm{VAR}[\mathrm{E}[W|Y_1]] \leq \mathrm{VAR}[\mathrm{E}[W|Y_1 Y_2]] \leq \mathrm{VAR}[W]$.

The interpretation of Proposition 2 is that as we add information to our estimate in the form of $Y = (Y_1)$ to $Y = (Y_1, Y_2)$, our measure for information increases. The lower bound represents the case of having zero information, and the upper bound represents the case of having full information.

Now, given a joint distribution p_{WY} between the edge weights W of the graph and the information Y that the agent receives, we can write the agent's average performance as

$$\begin{aligned}J(p_{WY}) &= \mathrm{E}\left[\min_{p\in P}\left\{p^T \mathrm{E}[W|Y]\right\}\right]\\ &= \mathrm{E}\left[\min_{p\in P}\left\{p^T \hat{W}\right\}\right] = J(p_{\hat{W}}).\end{aligned} \tag{4}$$

Notice that (4) only depends on $p_{\hat{W}}$.[2] We can also equivalently parameterize $\Gamma(C)$ by

$$\Gamma(C) = \left\{p_{\hat{W}}|\mathrm{VAR}[\hat{W}] \leq C \text{ and } \exists Y \text{ s.t. } \hat{W} \overset{\mathrm{d}}{=} \mathrm{E}[W|Y]\right\}. \tag{5}$$

F. Optional: Shortest Path Optimization Under a Mutual Information Bound

Why do we define a family of abstract sets $\{\Gamma(C)\}$ as our information bound and not simply use mutual information. In general, we want to apply bounds $\Gamma(C)$ that yield a nice relationship between C and $J(C)$. For instance, we will see that our variance bounds for information relate nicely to shortest path optimization.

We can relate our information bound to mutual information; however. Let

$$\Gamma'(C) = \left\{p_{WY}|I(W;Y) \leq I_{max}(C)\right\}$$

where $I_{max}(C) = \max_{p_{WY}\in\Gamma(C)}\{I(W;Y)\}$. Then $\Gamma(C) \subset \Gamma'(C)$ so that $J(\Gamma'(C)) \leq J(C)$.

In general, computing $I_{max}(C)$ is difficult, but so may be computing the performance under mutual information bounds directly. For instance, it is straightforward to see that the sets $\{\Gamma(C)\}$ that we are using for shortest path optimization relate to mutual information via a rate-distortion problem:

$$I_{max}(C) = \min_{\left\{p_{WY}|\mathrm{VAR}[W-\hat{W}]\leq \mathrm{VAR}[W]-C\right\}} I(W;Y).$$

In general, it is not trivial to solve this optimization, especially in the multivariable case.

G. Specializing to Gaussian Edge Weights

A particular subcase of interest to us is that of Gaussian edge weights. If $Y = W + N$ where W and N are independent, $W \sim N(\mu, \Lambda_W)$ with $\Lambda_W > 0$, and $N \sim N(0, \Lambda_N)$, then

$$\hat{W}(Y) = \Lambda_W(\Lambda_W + \Lambda_N)^{-1}(Y-\mu) + \mu.$$

Information optimization in this special case is equivalent to designing the distribution of the noise N, or, equivalently, its covariance matrix Λ_N. It is straightforward to show that designing a positive semidefinite Λ_N is equivalent to designing

[1] Note that Y is any side information, and so it can be taken as a tuple of side information as well.

[2] This is true for any linear objective, not just shortest path optimization.

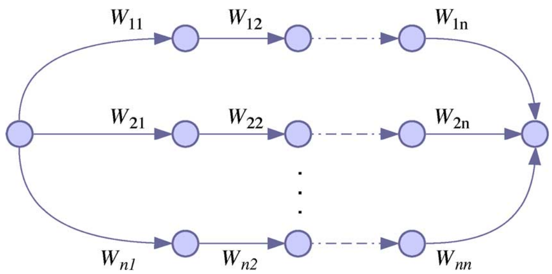

Fig. 1. Graph for Examples 1 and 2.

$\Lambda_{\hat{W}}$ (which is equal to $\Lambda_W(\Lambda_W+\Lambda_N)^{-1}\Lambda_W$) subject to certain constraints. Denote this new constraint set as $\Gamma_G(C)$:

$$\Gamma_G(C)=\left\{\Lambda_{\hat{W}}|0\preceq\Lambda_{\hat{W}}\preceq\Lambda_W \text{ and } \mathrm{Tr}(\Lambda_{\hat{W}})\le C\right\}. \quad (6)$$

Note that $\Gamma_G(C)$ is a convex set. It remains convex if we add additional convex constraints such as $\Lambda_{\hat{W}}\sim$ diag.

Finally, we define $J_G(C)=J(\Gamma_G(C))$.

H. *Impact of Information Selection: Comparative Examples*

We now briefly highlight the impact that different side information can have on performance.

Example 1: Let G be the graph in Fig. 1 having n disjoint paths $\{p_1,\dots,p_n\}$ from s to t with each path having n edges each. Let $W_{ij}\sim N(0,1)$ be the (random) weight of edge j on path i and assume the edge weights are independent. Consider a distribution p^s parameterized by covariance matrix $\Lambda^s\in\Gamma_G(n)$ where $\Lambda^s\sim$ diagonal, $\Lambda^s_{ee}=1$ if $e\in p_1$, and $\Lambda^s_{ee}=0$ otherwise. Essentially, the estimates $\hat{W}$ only contain information about the edges in p_1, meaning that $\hat{W}_e=W_e$ for $e\in p_1$ and $\hat{W}_e=0$ for $e\notin p_1$.

Under this side information, the average performance is

$$\begin{aligned}J(p^s)&=\mathrm{E}\left[\min_i\left\{\sum_j\hat{W}_{ij}\right\}\right]=\mathrm{E}\left[\min\left\{\sum_j W_{1j},0\right\}\right]\\&=\mathrm{E}\left[\min\left\{\sqrt{n}Z,0\right\}\right]=\sqrt{n}\mathrm{E}\left[\min\{Z,0\}\right]\\&=-\frac{1}{\sqrt{2\pi}}\sqrt{n}\end{aligned}$$

where $\sum_j W_{1j}\overset{\mathrm{d}}{=}\sqrt{n}Z$ with $Z\sim N(0,1)$.

Example 2: Let G be the graph in Fig. 1 and take the distribution p^p parameterized by a covariance matrix $\Lambda^p\in\Gamma_G(n)$ where $\Lambda^p\sim$ diagonal, $\Lambda^p_{ee}=1$ if e is the first edge of any path p_i, and $\Lambda^p_{ee}=0$ otherwise. Essentially, the estimates $\hat{W}$ only contain information about the first edge in each path, meaning that $\hat{W}_e=W_e$ if e is one of these links and $\hat{W}_e=0$ otherwise.

Under this side information, the average performance is

$$\begin{aligned}J(p^p)&=\mathrm{E}\left[\min_i\left\{\sum_j\hat{W}_{ij}\right\}\right]\\&=\mathrm{E}\left[\min\{W_{11}+0,W_{21}+0,\dots,W_{n1}+0\}\right]\\&=\mathrm{E}\left[\min\{Z_{11},Z_{21},\dots,Z_{n1}\}\right]\\&\ge-\sqrt{2\ln n}\end{aligned}$$

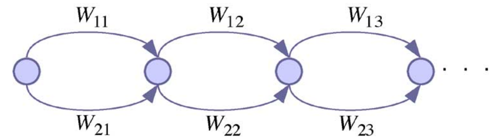

Fig. 2. Graph for Example 3.

where the last inequality is obtained by using Lemma 3 in [12].

There is a significant difference between the performances of the two examples. If we increase n, the average performance yielded from applying p^s outstrips that obtained using p^p quite substantially. This motivate our desire to optimize the information received by the agent.

The next example presents a topology for which the agent's performance significantly improves with capacity.

Example 3: Consider the graph in Fig. 2 with $W_{ij}\sim N(0,1)$ and independent and assume $C<(|E|/2)$. Choose any $\Lambda\in\Gamma_G(C)\bigcap\{\Lambda \text{ is diagonal}\}$, and denote $\lambda_e^2=\Lambda_{ee}$. We have

$$\begin{aligned}J(p_{\hat{W}})&=\sum_j\mathrm{E}\left[\min\{\hat{W}_{1j},\hat{W}_{2j}\}\right]\\&=\sum_j\mathrm{E}[\hat{W}_{2j}]+\mathrm{E}\left[\min\{\hat{W}_{1j}-\hat{W}_{2j},0\}\right]\\&=J(0)+\sum_j\min\left\{\sqrt{\lambda_{1j}^2+\lambda_{2j}^2}Z,0\right\}\\&=J(0)-\frac{1}{\sqrt{2\pi}}\sum_j\gamma_j\end{aligned}$$

where $\gamma_j=\sqrt{\lambda_{1j}^2+\lambda_{2j}^2}$ and $Z\sim N(0,1)$.

The optimal $\{\gamma_j^*\}$ (there are $|E|/2$ of them) are given by the solution to

$$\max_\gamma\left\{\sum_j\gamma_j\right\} \text{ subject to } \sum_j\gamma_j^2=C$$

yielding $\gamma_j^*=\sqrt{C/|E|/2}$. Thus,

$$J_G(C)\le J(0)-\frac{1}{\sqrt{2\pi}}\frac{|E|}{2}\sqrt{\frac{C}{|E|/2}}=J(0)-\frac{1}{2\sqrt{\pi}}\sqrt{|E|}\sqrt{C}.$$

Example 4: Finally, we consider a path having a single path from s to t. In this case, we get

$$J(C)=\mathrm{E}\left[\sum_e\hat{W}_e\right]=\sum_e\mathrm{E}[\hat{W}_e]=J(0)$$

where $J(0)$ is the performance under no information.

I. *Objective*

The goal of this paper is to develop practical algorithms for information optimization as well as analytic bounds for $J(C)$ that provide an intuitive relationship between capacity, topology, and performance.

III. Graph Characterization for Shortest Path Optimization

A challenge in information optimization is the evaluation of the objective $J(p_{\hat{W}})$, which is known to be #P-hard [13]. To overcome this difficulty, we will devote our analysis to developing meaningful upper and lower bounds for $J(p_{\hat{W}})$ that can be practically evaluated. In this section, we present a graph characterization for shortest path optimization that will allow us to derive such bounds. The characterization is based on a simplified description of the path polytope $\mathcal{P}$.

A. Projection Matrix for the Path Polytope

The first property of $\mathcal{P}$ that we examine will be critical to the development of our information optimization algorithms. First, define the path $\overline{p}$ of shortest average length as $\overline{p} = \arg\min_{p\in\mathcal{P}}\{p^T\mu\}$. Clearly, $J(0) = \overline{p}^T\mu$.

Proposition 3: $\mathcal{P} - \overline{p}$ lies in a strict subspace of $\Re^{|E|}$.

Proof: Because $0 \in \mathcal{P}$ and $\mathcal{P}$ is a polytope, we only need to show that it does not have volume in $\Re^{|E|}$.

Define the successor edge set E_v for the vertex v as $E_v = \{e|\mathrm{hd}(e) = v\}$. First, assume that $|E_s| > 1$. Define $H \in \Re^{|E_s|\times|E|}$ by

$$H_{(i,j)} = \begin{cases} 1, & i = j \in E_s \\ 0, & \text{otherwise.} \end{cases}$$

By the virtue of G being DAG and s being the unique start vertex, any path p in G must contain exactly one of the edges in E_s. Hence, the product HP is the set $\{p \in \Re^{|E_s|}|p_i = 1 \text{ for exactly one } i\}$, and therefore the product $H\mathcal{P}$ is the set of the corners of the simplex in $\Re^{|E_s|}$.

The simplex (the product $H\mathcal{P}$) does not have volume in $\Re^{|E_s|}$. Therefore, $\mathcal{P}$ does not contain a hypercube $\mathcal{C}$ of any size in $\Re^{|E|}$ because if it did, $H\mathcal{C}$ would be a set with volume in the simplex $H\mathcal{P}$. Hence, $\mathcal{P}$ has no volume.

Now, assume that $|E_s| \leq 1$. If its equals zero, the claim is obviously true. If it equals one, select the first vertex v "after" s with $|E_v| > 1$ and apply the proof above to E_v to show that $\mathcal{P}$ has no volume. If $|E_v| = 1$ for all v (the only remaining case), then one can easily see that there is only one path in G, the vector $[1\ 1\ \ldots\ 1]$, which is a single point in $\Re^{vertE|}$ and, thus, has no volume. ■

Let $S_{\mathcal{P}}$ be the smallest subspace containing $\mathcal{P}-\overline{p}$, and let $H_{\mathcal{P}}$ be the projection matrix for $S_{\mathcal{P}}$. We can compute $H_{\mathcal{P}}$ in polynomial time, but the details of the computation are not critical to the developments of this paper, so we save the details for the Appendix. The following theorem formally states our claim.

Proposition 4: $H_{\mathcal{P}}$ can be computed in polynomial time.

B. Outer Spheric Approximation

Now we provide a simplified description of the boundary for $\mathcal{P}$ using a low-complexity outer approximation, specifically a sphere. In general, finding the minimal-radius sphere containing a polytope is computationally hard [14], but in the case of the path polytope, it can be computed quite efficiently.

Proposition 5: The minimal-radius sphere $B(r_o^*, c_o^*)$ containing $\mathcal{P}$ is given by the solution to the convex quadratic optimization

$$\begin{aligned} &\min_{r,c\in\Re^{|E|},\overline{J}:V\to\Re} \quad r^2 \text{ subject to} \\ &r^2 \geq \overline{J}(s) + \|c\|^2 \\ &\overline{J}(v) \geq \max_{e|\mathrm{hd}(e)=v} \left\{\overline{J}\left(\mathrm{tl}(e)\right) + (1-2c_e)\right\} \\ &\overline{J}(t) = 0. \end{aligned} \tag{7}$$

Proof: Because the extreme points of $\mathcal{P}$ are the paths p of the graph, a necessary and sufficient condition for a sphere with radius r and center c to contain $\mathcal{P}$ is

$$\|p - c\|^2 \leq r^2 \Leftrightarrow p^Tp - 2p^Tc + c^Tc \leq r^2.$$

Since p is a 0–1 vector, $p^Tp = 1^Tp$, so we have the equivalent inequality

$$\begin{aligned} r^2 &\geq p^T(1-2c) + \|c\|^2 \text{ for all } p \in P \\ \Leftrightarrow r^2 &\geq \max_{p\in P}\left\{p^T(1-2c)\right\} + \|c\|^2 \\ \Leftrightarrow r^2 &\geq \max_{p\in\mathcal{P}}\left\{p^T(1-2c)\right\} + \|c\|^2 \end{aligned}$$

where the last two lines differ in the use of P versus its convex hull $\mathcal{P}$ (which is possible since it is a linear optimization). The expression $\max_{p\in\mathcal{P}}\{p^T(1-2c)\}$ is the length of the longest path in an acyclic graph when the edge weights are given by $(1-2c)$. It is straightforward to see that the optimizing $\overline{J}(s)$ is this length. ■

Remark 2: Analytic bounds for r_o^* can be computed by first choosing a (possibly non-optimal) center c, computing the length of the longest path using the edge weight vector $(1-2c)$, and then computing a radius bound. We will use this strategy in some examples later in the paper.

Proposition 6: $\mathcal{P} - c_o^* \subset \mathcal{S}_{\mathcal{P}}$.

Proof: Suppose the minimum radius ball $B = B(r_o^*, c_o^*)$ has its center c_o^* not in $\mathcal{P}$. Then there is a hyperplane strictly separating $\mathcal{P}$ and c_o^*. This hyperplane also defines a closed half-space L that contains $\mathcal{P}$ but not c_o^*. The diameter of $L\cap B < 2r_o^*$, since the existence of points $p_1, p_2 \in L\cap B$ with $\|p_1 - p_2\|_2 = 2r_o^*$ would imply that $c_o^* = (1/2)(p_1+p_2)$ is in $L\cap B$. Therefore, $L\cap B$ can be enclosed in a ball of radius less than r_o^*, implying that $\mathcal{P}$ can be enclosed in a ball with radius less than r_o^*. This is a contradiction. ■

C. Inner Spheric Approximation

Efficient algorithms for generating inner spheric (as well as ellipsoidal) approximations to a polytope are well-known [14], and we do not reproduce the results in this paper. We do note, however, than such algorithms assume that the set to be approximated has volume, which $\mathcal{P}$ does not. A way around this problem is to simply compute the inner approximation strictly within the affine subspace containing $\mathcal{P}$. The details of the computation are not important to the developments of this paper since we will not leverage inner spheric approximations beyond some basic statements that are nearly identical to the case of outer spheric approximations.

Similar to the case of an outer spheric approximation, we let r_i^* and c_i^* respectively be the radius and center of the maximal inner sphere contained in $\mathcal{P}$.

IV. Information Optimization and Analytic Performance Bounds Using Graph Reductions

In this section, we derive expressions for information optimizations, analytic performance bounds, and a numeric technique for bounding $J(C)$.

A. Information Optimizations Via Upper and Lower Bound Optimization

We begin with a derivation of upper and lower bounds for $J(C)$ that can be used for information optimization. While these bounds are not practical to optimize, they do provide useful insight into the relationship between information and performance.

Lemma 1:

$$(c_o^*)^T \mu - r_o^* \max_{p_{\hat{W}} \in \Gamma(C)} \left\{ \mathrm{E}\left[\|H_{\mathcal{P}} \hat{W}\| \right] \right\} \leq J(C) \leq (c_i^*)^T \mu - r_i^* \min_{p_{\hat{W}} \in \Gamma(C)} \left\{ \mathrm{E}\left[\|H_{\mathcal{P}} \hat{W}\| \right] \right\}.$$

Proof: We start with the lower bound. Since $c_o^* \in \mathcal{P}$, $H_{\mathcal{P}}(\mathcal{P} - c_o^*) = \mathcal{P} - c_o^*$. Therefore,

$$\begin{aligned} J(p_{\hat{W}}) &= \mathrm{E}\left[\min_{p \in \mathcal{P} - c_o^*} \left\{ (p + c_o^*)^T \hat{W} \right\} \right] \\ &= \mathrm{E}\left[\min_{p \in H_{\mathcal{P}}(\mathcal{P} - c_o^*)} \{ p^T \hat{W} \} \right] + (c_o^*)^T \mu \\ &= \mathrm{E}\left[\min_{p \in (\mathcal{P} - c_o^*)} \left\{ (H_{\mathcal{P}} p)^T \hat{W} \right\} \right] + (c_o^*)^T \mu \\ &\geq \mathrm{E}\left[\min_{p \in B(r_o^*, 0)} \{ p^T H_{\mathcal{P}} \hat{W} \} \right] + (c_o^*)^T \mu \\ &= -r_o^* \mathrm{E}\left[\left\| H_{\mathcal{P}}(\hat{W}) \right\| \right] + (c_o^*)^T \mu. \end{aligned}$$

The upper bound similarly follows. ■

An interpretation of Theorem 1 is that we should optimize side information by concentrating the energy of the estimate to the subspace $\mathcal{S}_{\mathcal{P}}$. The component of $\hat{W}$ normal to that subspace (specifically, $(I - H_{\mathcal{P}})\hat{W}$) is lost in the projection. To this end, we term $\mathcal{S}_{\mathcal{P}}$ the *actionable* subspace, and we term the orthogonal subspace the *inactionable* subspace.

What does the inactionable component of the information represent. It is the amount of length in $\hat{W}$ common to all paths. Thus, it only aids in estimating the actual length of the paths. The actionable component $H_{\mathcal{P}} \hat{W}$ contains all of the information for path selection. We will specialize these bounds in Section III to the case of Gaussian edge weights to give practical information optimization algorithms.

We can derive a different information optimization for $J(C)$ by choosing a simpler strategy: optimize information over only two paths of a graph. Geometrically, we achieve this by reducing $\mathcal{P}$ to a line (i.e., by connecting two points in $\mathcal{P}$) and removing the remainder of the polytope.

Lemma 2: For any $\hat{p} \in \mathcal{P}$, an upper bound for $J(C)$ is

$$J(C) \leq J(0) + \min_{p_{\hat{W}} \in \Gamma(C)} \left\{ \mathrm{E}\left[\min\left\{ 0, (\hat{p} - \overline{p})^T \hat{W} \right\} \right] \right\}.$$

Proof:

$$\begin{aligned} J(p_{\hat{W}}) &= \mathrm{E}\left[\min_{p \in \mathcal{P}} \{ p^T \hat{W} \} \right] \leq \mathrm{E}\left[\min_{p \in \{\overline{p}, \hat{p}\}} \{ p^T \hat{W} \} \right] \\ &= \mathrm{E}\left[\min\{ \overline{p}^T \hat{W}, \hat{p}^T \hat{W} \} \right] \\ &= \mathrm{E}[\overline{p}^T \hat{W}] + \mathrm{E}\left[\min\left\{ 0, (\hat{p} - \overline{p})^T \hat{W} \right\} \right]. \end{aligned}$$

Because the bound is true for all $p_{\hat{W}}$, it is true for the minimizing distribution. ■

Because $\hat{p} - \overline{p}$ lies in the actionable subspace $\mathcal{S}_{\mathcal{P}}$, the optimizing distribution $p_{\hat{W}}$ will concentrate $\hat{W}$ to the actionable subspace as well. We will see that this strategy will arise again later in this paper. It is also in qualitative agreement with Examples 1 and 2 where it was shown that concentrating information energy along a single path of a graph is superior to spreading it among many paths.

B. Analytic Performance Lower Bounds

Using Lemma 1, we can derive an analytic performance bound for $J(C)$ that incorporates the capacity C and graph radius r_o^*. We begin two useful background results.

Proposition 7: For any function $f : P \to \Re$ and any set $\overline{\mathcal{P}} \supset \mathcal{P}$

$$J(p_{\hat{W}}) \geq \min_{p \in \mathcal{P}} \{ f(p) \} + \mathrm{E}\left[\min_{p \in \overline{\mathcal{P}}} \left\{ p^T \hat{W} - f(p) \right\} \right].$$

Proof: The proof follows immediately from the fact that $\min_x\{a(x) + b(x)\} \geq \min_x\{a(x)\} + \min_x\{b(x)\}$. ■

Remark 3: Proposition 7 provides a generalization of the approach taken in [15] to generate a low-complexity optimization for bounding the mean of the minimum order statistic. Let $\mathcal{P}$ be the simplex in $\Re^n$ and let $W = (W_1, \ldots W_n)$ be a random vector in $\Re^n$. The minimum order statistic is given by $\min_i\{W_i\} = \min_{x \in \mathcal{X}}\{x^T W\}$. We can derive the bound in [15] by setting $f(x) = x^T z$ for some vector $z \in \Re^n$, setting $\overline{\mathcal{P}}$ to the unit cube, and then maximizing over z. A similar approach is taken in [16] for obtaining an analytic lower bound for $J(C)$.

Proposition 8:

$$J(C) \geq J(0) - r_o^* \max_{p_{\hat{W}} \in \Gamma(C)} \left\{ \mathrm{E}\left[\left\| H_{\mathcal{P}}(\hat{W} - \mu) \right\| \right] \right\}.$$

Proof: Applying Proposition 7 with $f(p) = p^T \mu$ and $\overline{\mathcal{P}} = \mathcal{P}$ yields

$$\begin{aligned} J(p_{\hat{W}}) &\geq \min_{p \in \mathcal{P}} \{ p^T \mu \} + \mathrm{E}\left[\min_{p \in \mathcal{P}} \left\{ p^T (\hat{W} - \mu) \right\} \right] \\ &= J(0) + \mathrm{E}\left[\min_{p \in \mathcal{P}} \left\{ p^T (\hat{W} - \mu) \right\} \right]. \end{aligned}$$

We can now apply Lemma 1 using $\hat{W} - \mu$ in place of $\hat{W}$ yielding

$$J(C) \geq J(0) - r_o^* \max_{p_{\hat{W}} \in \Gamma(C)} \left\{ \mathrm{E}\left[\left\| H_{\mathcal{P}}(\hat{W} - \mu) \right\| \right] \right\}.$$

■

We now provide the analytic performance bound.

Theorem 1: $J(C) \geq J(0) - r_o^* \sqrt{C}$.

Proof: By $\sigma_{\max}(H_{\mathcal{P}}) = 1$ and Jensen's Inequality

$$\mathrm{E}\left[\|H_{\mathcal{P}}\hat{W}\|\right] \leq \sqrt{\mathrm{E}\left[\|H_{\mathcal{P}}\hat{W}\|^2\right]} \leq \sqrt{\mathrm{E}\left[\|\hat{W}\|^2\right]} \leq \sqrt{C}$$

for all distributions $p_{\hat{W}} \in \Gamma(C)$. Simply apply this inequality to Corollary 8. ∎

Finally, the corollary below gives the maximum rate of performance improvement one can get for any graph topology.

Corollary 1: A proportionally tight lower bound for $J(C)$ over all graph topologies is $J(C) \geq J(0) - (1/2)\sqrt{|E|}\sqrt{C}$.

Proof: Substituting $c = [(1/2)\ldots(1/2)]^T$ in the optimization in Proposition 5 yields $r_o^* \leq (1/2)\sqrt{|E|}$. Proportional tightness is proven in Example 3. ∎

Remark 4: The bound in Corollary 1 tells us that there exists a graph topology (the one in Example 3) such that performance unboundedly improves as the graph grows, even if $C > 0$ is held fixed.

Remark 5: The bound in Corollary 1 also appears in [16] and [8] using different methods. In [16], it is obtained by bounding $\mathcal{P}$ with the unit cube. In [8], it is obtained using convex majorization of RVs. The methods used in these other papers cannot be improved further to include graph topology information.

C. A Numeric Approach to Computing Performance Bounds

We now present a mathematical programming approach to computing bounds for $J(C)$. The approach is based on the results in [10] for bounding general stochastic optimizations. First, consider the following optimization:

$$\begin{aligned} J_{opt}(C) = \min_{p_{\hat{W}}} & \left\{ \mathrm{E}\left[\min_{p \in P}\{p^T \hat{W}\}\right]\right\} \text{ subject to} \\ & \mathrm{E}[\hat{W}] = \mu, \mathrm{VAR}\,[\hat{W}] \leq C \\ & 0 \leq \mathrm{VAR}\,[\hat{W}_e] \leq \sigma_e^2. \end{aligned} \tag{8}$$

Clearly, $J(C) \geq J_{opt}(C)$ since the constraint set of the optimization is a superset of $\Gamma(C)$.[3]

We can solve (8) by solving its dual, and there are conditions under which there is no duality gap [11]. However, the dual will have a constraint for each path in the graph, and so it may be impractical to solve for even a moderately sized graph. An efficient approach to tackling optimizations similar to (8) is presented in [10], but it requires that the primal only contains constraints on the individual edge weight moments, which the capacity constraint $\mathrm{VAR}[\hat{W}] \leq C$ clearly disobeys. [17] extends [10] to the case of having constraints on the non-diagonal components of the covariance matrix, but our constraint is of a different form. Nonetheless, we can adapt the approach in [10] to handle our capacity constraint. We present this extension as a corollary to the main result Corollary 3.2 in [10]. The proof is contained in the Appendix.

[3]Notice that there are no higher order or cross moment constraints (such as a covariance constraint) on $p_{\hat{W}}$ in terms of p_W in (8). Hence, we cannot guarantee that there is an RV Y such that an RV $\hat{W}$ from this constraint set satisfies $\hat{W} = \mathrm{E}[W|Y]$.

Corollary 2 (to Corollary 3.2 in [10]):

$$\begin{aligned} J_{opt}(C) = \min_{\{H_e\},\{\lambda_e\}} & \left\{\sum_e \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \cdot H_e \right\} \text{ subject to} \\ & H_e \leq 0, H_e \geq -\begin{bmatrix} \lambda_e^2 + \mu_e^2 & \mu_e \\ \mu_e & 1 \end{bmatrix} \\ & \sum_e \left(\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \cdot H_e \right) v_e \in \mathcal{P} \\ & 0 \leq \lambda_e^2 \leq \sigma_e^2, \sum_e \lambda_e^2 \leq C \end{aligned} \tag{9}$$

where v_e is the elementary basis vector with the e^{th} component equal to 1.

A problem with (8) and (9) for information optimization is that the optimizing distribution $p_{\hat{W}}^*$ may not be possible to achieve for a given edge weight distribution p_W. Therefore, it cannot be used to derive information optimization algorithms since there may be no side information distribution p_{WY} that can yield $p_{\hat{W}}$. Also noteworthy is that although (9) can provide performance bounds that use the graph's topology, its computational nature does not provide any intuition for how topology or capacity impact performance.

V. Information Optimization and Performance Bounds in the Gaussian Case

A. Information Optimizations Via Upper and Lower Bound Optimization

In the Gaussian case, we can provide very useful upper and lower bounds for $J(p_{\hat{W}})$ that are amenable to information optimization. We begin with a useful background result.

Lemma 3: Let $Z \sim N(0, I)$. The optimization

$$\max_{\Lambda_{\hat{W}} \in \Gamma_G(C)} \left\{ \mathrm{E}\left[\|H_{\mathcal{P}}\sqrt{\Lambda_{\hat{W}}} Z\|\right]\right\}$$

is equivalent to the convex optimization

$$\max_{\Lambda_{\hat{W}} \in \mathcal{L}} \left\{ \mathrm{E}\left[\|\sqrt{\Lambda_{\hat{W}}} Z\|\right]\right\}$$

where $\mathcal{L} = \Gamma_G(C) \bigcap \{\Lambda_{\hat{W}} | (H_{\mathcal{P}} - I)\Lambda_{\hat{W}} = 0\}$.

Proof: We have

$$\begin{aligned} & \max_{\Lambda_{\hat{W}} \in \Gamma_G(C)} \left\{ \mathrm{E}\left[\|H_{\mathcal{P}}\sqrt{\Lambda_{\hat{W}}} Z\|\right]\right\} \\ & \quad = \max_{\Lambda_{\hat{W}} \in \Gamma_G(C), \sqrt{\Lambda_{\hat{W}}} = H_{\mathcal{P}}\sqrt{\Lambda_{\hat{W}}}} \left\{ \mathrm{E}\left[\|\sqrt{\Lambda_{\hat{W}}} Z\|\right]\right\} \end{aligned}$$

where the equality comes from the fact that $H_{\mathcal{P}} H_{\mathcal{P}} = H_{\mathcal{P}}$.

Since $\mathrm{Range}(\Lambda_{\hat{W}}) = \mathrm{Range}(\sqrt{\Lambda_{\hat{W}}})$, we have the following equivalences for the constraint $\sqrt{\Lambda_{\hat{W}}} = H_{\mathcal{P}}\sqrt{\Lambda_{\hat{W}}}$:

$$\begin{aligned} & \sqrt{\Lambda_{\hat{W}}} = H_{\mathcal{P}}\sqrt{\Lambda_{\hat{W}}} \Leftrightarrow \mathrm{Range}(\sqrt{\Lambda_{\hat{W}}}) = \mathcal{S}_{\mathcal{P}} \\ & \Leftrightarrow \mathrm{Range}(\Lambda_{\hat{W}}) = \mathcal{S}_{\mathcal{P}} \Leftrightarrow \Lambda_{\hat{W}} = H_{\mathcal{P}}\Lambda_{\hat{W}} \\ & \Leftrightarrow (H_{\mathcal{P}} - I)\Lambda_{\hat{W}} = 0. \end{aligned}$$

Clearly, this constraint is convex over $\Lambda_{\hat{W}}$. Therefore, $\mathcal{L}$ is convex.

Now we prove the objective is concave. The function $f(\Lambda_{\hat{W}}) = z^T\Lambda_{\hat{W}}z = \|\sqrt{\Lambda_{\hat{W}}}z\|^2$ is linear over $\Lambda_{\hat{W}} \geq 0$. Therefore, $\sqrt{f(\Lambda_{\hat{W}})} = \|\sqrt{\Lambda_{\hat{W}}}z\|$ is concave for $\Lambda_{\hat{W}} \geq 0$. By linearity of the expected value operator, the objective is concave. ■

We now present the main information optimization result in the Gaussian case.

Theorem 2: If the mean lengths of all paths are less than a constant K (formally, $p^T\mu \leq K$ for all paths p), then

$$J(0) - r_o^* \max_{\Lambda_{\hat{W}}\in\mathcal{L}} \left\{\mathrm{E}\left[\|\sqrt{\Lambda_{\hat{W}}}Z\|\right]\right\} \leq J_G(C) \leq K - r_i^* \max_{\Lambda_{\hat{W}}\in\mathcal{L}} \left\{\mathrm{E}\left[\|\sqrt{\Lambda_{\hat{W}}}Z\|\right]\right\} \quad (10)$$

where $Z \sim N(0, I)$ and $\mathcal{L} = \Gamma_G(C) \bigcap \{\Lambda_{\hat{W}} | (H_\mathcal{P} - I)\Lambda_{\hat{W}} = 0\}$.

Proof: We start by proving the lower bound. By Proposition 8,

$$J_G(C) \geq J(0) - r_o^* \max_{p_{WY}\in\Gamma_G(C)} \left\{\mathrm{E}\left[\left\|H_\mathcal{P}(\hat{W} - \mu)\right\|\right]\right\}.$$

By our parameterization $\hat{W} = \sqrt{\Lambda_{\hat{W}}}Z + \mu$, we have the equivalent expression for the maximization term:

$$\max_{\Lambda_{\hat{W}}\in\Gamma_G(C)} \left\{\mathrm{E}\left[\|H_\mathcal{P}\sqrt{\Lambda_{\hat{W}}}Z\|\right]\right\}.$$

The remainder of the proof for the lower bound follows from Lemma 3.

Now we turn our attention to the upper bound. If $p^T\mu \leq K$, then

$$\begin{aligned} J_G(p_{\hat{W}}) &= \mathrm{E}\left[\min_p\{p^T\hat{W}\}\right] \\ &\leq \mathrm{E}\left[\min_p\left\{p^T\hat{W} + (K - p^T\mu)\right\}\right] \\ &= \mathrm{E}\left[\min_p\left\{p^T(\hat{W} - \mu)\right\}\right] + K \\ &\leq K - r_i^* \max_{p_{WY}\in\Gamma(C)} \left\{\mathrm{E}\left[\left\|H_\mathcal{P}(\hat{W} - \mu)\right\|\right]\right\} \end{aligned}$$

where the last inequality follows from Lemma 1. The remainder of the proof is similar to that of the lower bound. ■

B. Special Case Analytic Solutions for Information Optimization

Theorem 2 provides convex optimizations for information optimization. Under certain conditions, we can derive analytic expressions for the optimal bounds as well as the optimizing covariance matrix $\Lambda_{\hat{W}}^*$. We begin with the following lemma.

Lemma 4: If C is sufficiently small, then

$$\max_{\Lambda_{\hat{W}}\in\Gamma_G(C)} \left\{\mathrm{E}\left[\|H_\mathcal{P}\sqrt{\Lambda_{\hat{W}}}Z\|\right]\right\} = \sqrt{\frac{2}{\pi}}\sqrt{C}$$

where an optimizing $\Lambda_{\hat{W}}^*$ is $C/\|p_1 - p_2\|^2(p_1 - p_2)(p_1 - p_2)^T$ for two chosen paths $p_1, p_2 \in P$.

Proof: By $\sigma_{\max}(H_\mathcal{P}) = 1$, $\|H_\mathcal{P}\Lambda_{\hat{W}}^{1/2}Z\| \leq \|\Lambda_{\hat{W}}^{1/2}Z\|$. Therefore, $\Lambda_{\hat{W}}^*$ must satisfy $H_\mathcal{P}\sqrt{\Lambda_{\hat{W}}^*}Z = \sqrt{\Lambda_{\hat{W}}^*}Z$ to maximize the objective. We can parameterize the set of all feasible $\Lambda_{\hat{W}} \in \Gamma_G(C)$ satisfying this constraint as follows. Let $m = \dim(\mathcal{S}_\mathcal{P})$. Set $\Lambda_{\hat{W}} = U\Sigma U^T$ where the unitary matrix U has the form

$$U = [x_1 \cdots x_m\, y_{m+1} \cdots y_{|E|}]$$

with $\{x_i\}$ and $\{y_i\}$ form an orthonormal bases for $\mathcal{S}_\mathcal{P}$ and $\mathcal{S}_\mathcal{P}^\perp$ respectively, Σ is a diagonal matrix with $\Sigma_{ii} = C_i$ for $i \leq m$, $\Sigma_{ii} = 0$ for $i > m$, and $C_i \geq 0$ satisfy $\sum_i C_i = C$. Optimizing $\Lambda_{\hat{W}}$ is now equivalent to optimizing over the basis $\{x_i\}$ and scalars $\{C_i\}$.

Substituting this parameterization for $\Lambda_{\hat{W}}$ yields the optimization

$$\max_{\{x_i\},\{C_i\}} \left\{\mathrm{E}\left[\left\|\sum_{i=1}^m C_i^{\frac{1}{2}} x_i x_i^T Z\right\|\right]\right\} = \max_{\{x_i\},\{C_i\}} \left\{\mathrm{E}\left[\sqrt{\sum_{i=1}^m C_i\left(x_i^T Z\right)^2}\right]\right\}$$

where equality follows from the fact that the $\{x_i\}$ are an orthonormal basis.

By the concavity of $\sqrt{\cdot}$ and the fact that the vector $(C_1/C, \ldots, C_m/C)$ is on a simplex, we have

$$\begin{aligned} &= \max_{\{x_i\},\{C_i\}} \left\{\sqrt{C}\mathrm{E}\left[\sqrt{\sum_{i=1}^m \left(\frac{C_i}{C}\right)\left(x_i^T Z\right)^2}\right]\right\} \\ &\leq \max_{\{x_i\},\{C_i\}} \left\{\sqrt{C}\sum_{i=1}^m \left(\frac{C_i}{C}\right)\mathrm{E}\left[\sqrt{\left(x_i^T Z\right)^2}\right]\right\} \\ &= \max_{\{x_i\},\{C_i\}} \left\{\sqrt{C}\sum_{i=1}^m \left(\frac{C_i}{C}\right)K\right\} \\ &= \max_{\{x_i\},\{C_i\}} \{\sqrt{C}K\} = \sqrt{C}K \end{aligned}$$

where $K = \mathrm{E}[\sqrt{(n^T Z)^2}]$ for any unit vector n (by symmetry, any unit vector n yields the same number, and hence it is the same for each x_i). If we choose $n = [1\ 0 \cdots 0]^T$, we get $K = \mathrm{E}[|Z_1|]$ where $|Z_1|$ has the folded normal distribution. Hence, $K = \sqrt{2/\pi}$.

The upper bound is achieved if we choose any orthonormal basis $\{x_i\}$ and set $C_i = C$ for some i and $C_j = 0$ for all $j \neq i$. Therefore, for some i, set $x_i = p_1 - p_2/\|p_1 - p_2\|$ since then $x_i \in \mathcal{S}_\mathcal{P}$, set $C_i = C$, and set the remaining x_j's to some orthogonal basis with $C_j = 0$. Note that we require $\Lambda_{\hat{W}} \preceq \Lambda_W$. We assume C is sufficiently small in the theorem statement so that this condition is satisfied. ■

Remark 6: Note that if we were to include additional constraints such as $\Lambda_{\hat{W}} \sim$ diagonal, we may not be able to derive an analytic solution for the optimizing $\Lambda_{\hat{W}}^*$. In this case, we would have to numerically solve the convex optimization.

Theorem 3: Under the assumptions of Lemma 4 and if $p^T\mu \leq K$ for all paths p, then

$$J(0) - r_o^*\sqrt{\frac{2}{\pi}}\sqrt{C} \leq J_G(C) \leq K - r_i^*\sqrt{\frac{2}{\pi}}\sqrt{C}$$

An interesting property of the optimizing $\Lambda_{\hat{W}}^*$ in Lemma 4 is that it concentrates the information to two paths of the graph over the edges for they do not intersect (the $p_1 - p_2$ expression). This is in qualitative agreement with Examples 1 and 2 where it

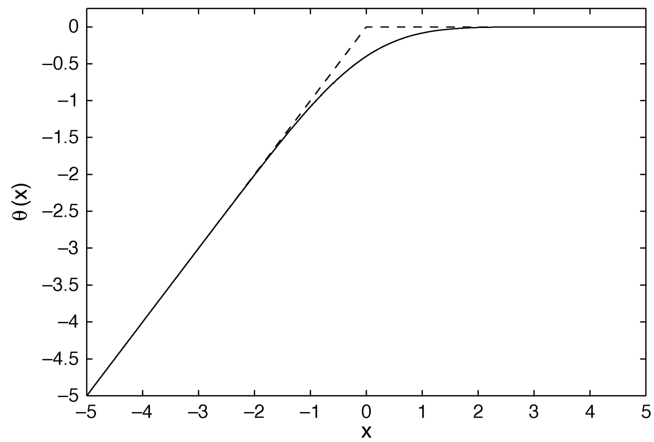

Fig. 3. $\Theta(x)$ (solid line) compared to $\min\{x, 0\}$ (dashed line).

was shown that concentrating information energy along a single path of a graph is superior to spreading it among many paths. Hence, we can consider deriving a new upper bound for $J_G(C)$ simply by concentrating information in this fashion.

First, let $\Theta(c) = \mathrm{E}[\min\{X, c\}]$ where $X \sim N(0,1)$. Clearly, $\Theta(c) \leq c$. Essentially, $\Theta(c)$ seeks to determine how much X is less than c on average (see Fig. 3). One can easily show that $\Theta(0) = -1/\sqrt{2\pi}$.

Theorem 4: Assume independent edge weights. For any $\hat{p} \in \mathcal{P}$, an upper bound for $J_G(C)$ is

$$J_G(C) \leq J(0) + \sqrt{\tilde{C}}\Theta\left(\frac{(\hat{p}-\overline{p})^T\mu}{\sqrt{\tilde{C}}}\right)$$

where $\tilde{C} = \min\{C, \mathrm{VAR}[(\hat{p}-\overline{p})^T W]\}$.

Proof: Under independence assumptions, it is easy to verify that

$$\mathrm{VAR}\left[(\hat{p}-\overline{p})^T\hat{W}\right] \leq \min\left\{C, \mathrm{VAR}\left[(\hat{p}-\overline{p})^T W\right]\right\} = \tilde{C}.$$

It is also easy to construct a distribution $p_{\hat{W}}$ that corresponds to a diagonal $\Lambda_{\hat{W}} \in \Gamma_G(C)$ that achieves this bound. Therefore,

$$(\hat{p}-\overline{p})^T\hat{W} \stackrel{\mathrm{d}}{=} \sqrt{\tilde{C}}Z + (\hat{p}-\overline{p})^T\mu.$$

where $Z \sim N(0,1)$.

Beginning with (an adaptation of) Lemma 2, we have

$$\begin{aligned} J_G(C) &\leq J(0) + \min_{p_{\hat{W}\in\Gamma_G(\tilde{C})}}\left\{\mathrm{E}\left[\min\left\{0, (\hat{p}-\overline{p})^T\hat{W}\right\}\right]\right\} \\ &= J(0) + \mathrm{E}\left[\min\left\{\sqrt{\tilde{C}}Z + (\hat{p}-\overline{p})^T\mu, 0\right\}\right] \\ &= J(0) + \mathrm{E}[\sqrt{\tilde{C}}Z] + \mathrm{E}\left[\min\left\{(\hat{p}-\overline{p})^T\mu, -\sqrt{\tilde{C}}Z\right\}\right] \\ &= J(0) + 0 + \sqrt{\tilde{C}}\mathrm{E}\left[\min\left\{\frac{(\hat{p}-\overline{p})^T\mu}{\sqrt{\tilde{C}}}, Z\right\}\right]. \end{aligned}$$

■

Corollary 3: Assume the conditions of Theorem 4 and further assume that $\hat{p}^T\mu = \overline{p}^T\mu$. An upper bound for $J_G(C)$ is

$$J_G(C) \leq J(0) - \frac{1}{\sqrt{2\pi}}\sqrt{\tilde{C}}$$

where $\tilde{C} = \min C, \mathrm{VAR}[(\hat{p}-\overline{p})^T W]$.

Remark 7: There are two key differences between the upper bounds in Corollary 3 and Theorem 3. First is the constant. By restricting the polytope $\mathcal{P}$ to a line, Corollary 3 only concerns itself with the average-optimal length $J(0)$ rather than an upper bound for all path lengths. Second is the coefficient in front of the $\sqrt{C}$ term: $1/\sqrt{2\pi}$ in Corollary 3 and $r_i^*\sqrt{2/\pi}$ in Theorem 3. The inner radius r_i^* can actually act as an impediment on the bound since $r_i^* \to 0$ for certain path polytopes (such as the simplex) as the dimension increases, making the bound potentially over-conservative.

VI. Examples

We begin with two analytic examples that bound the value of side information for certain graph topologies and then conclude the section with two numeric examples.

Example 5: In this example, we compute an analytic performance bound for a binary tree with L levels. We first choose a center c for the outer sphere approximation by selecting a center point c_e for each edge e. A binary tree consists of levels of edges, an edge's level being defined as its distance from the root vertex. We select c according to the scheme $c_e = 2^{-l_e}$ where l_e is the level of edge e. Because level l has 2^l edges in it, we have 2^l edges e with $c_e = 2^{-l}$.

We compute a bound for r_o^* using this center as follows:

$$\begin{aligned} (r_o^*)^2 &\leq \max_p\left\{p^T(1-2c)\right\} + \|c\|^2 \\ &= \left(1 - 2\sum_{l=1}^{L}\frac{1}{2^l}\right) + \left(\sum_l\sum_{e\in\text{layer } l} 2^l\frac{1}{2^l}\right) \\ &= \frac{1}{2^{L-1}} - 1 + L \\ &\leq L. \end{aligned}$$

Therefore, the average performance under capacity C is lower bounded by $J(C) \geq J(0) - \sqrt{L}\sqrt{C}$.

Example 6: We now compute an analytic performance bound for a complete graph where the start vertex s and terminating vertex t are chosen arbitrarily. Of course, since a complete graph is undirected, a direct application of (7) is not practical since it requires the computation the longest path in an undirected graph with non-negative edge weights. However, we can compute a suboptimal analytic lower bound fairly easily.

For a selection of s and t in the complete graph, we divide the edges of the graph into classes of edges and apply the same center point to each edge within the same class. The classes are:

- $\mathcal{A} = \{$the edge connecting s and $t\}$;
- $\mathcal{B} = \{$edges connecting s to vertices $u \neq t\}$;
- $\mathcal{C} = \{$edges connecting $v \neq s$ to $t\}$;
- and $\mathcal{D} = \{$edges connecting $v \neq s$ to $u \neq t\}$.

We can show that $|\mathcal{A}| = 1$, $|\mathcal{B}| = |\mathcal{C}| = |V| - 2$, and $|\mathcal{D}| = (|V|-2)(|V|-3)/2$.

We assign the center points as follows:

- $e \in \mathcal{A} \Rightarrow c_e = a$;
- $e \in \mathcal{B} \Rightarrow c_e = b$;
- $e \in \mathcal{C} \Rightarrow c_e = c$;
- and $e \in \mathcal{D} \Rightarrow c_e = d$.

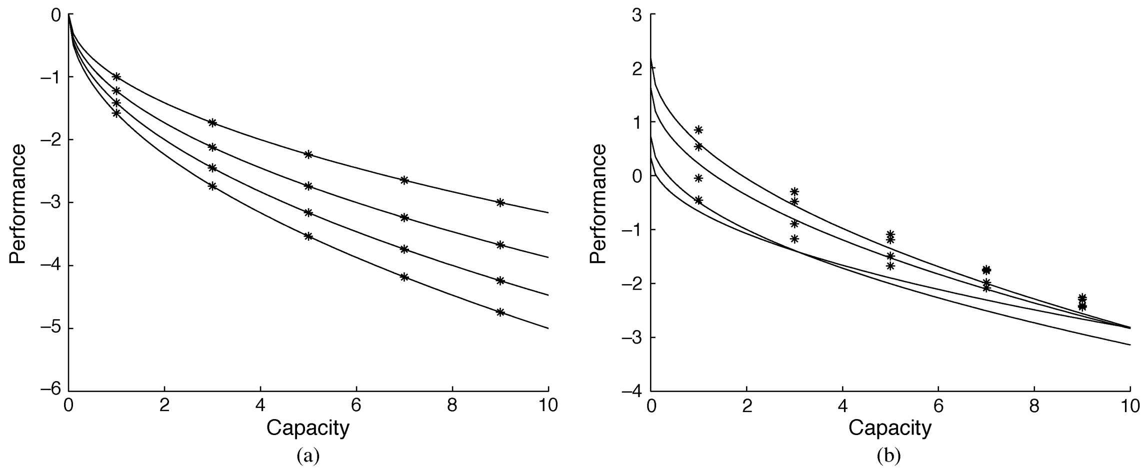

Fig. 4. Analytic bound of Theorem 1 compared to the optimization-based bound of Corollary 2 for a two-path graph. The solid curves are the analytic bound performances, and the asterisks are the optimization-based bound performances. Each curve represents a graph with an different number of links per path. (a) $\mu = 0$ case. (b) Random μ case.

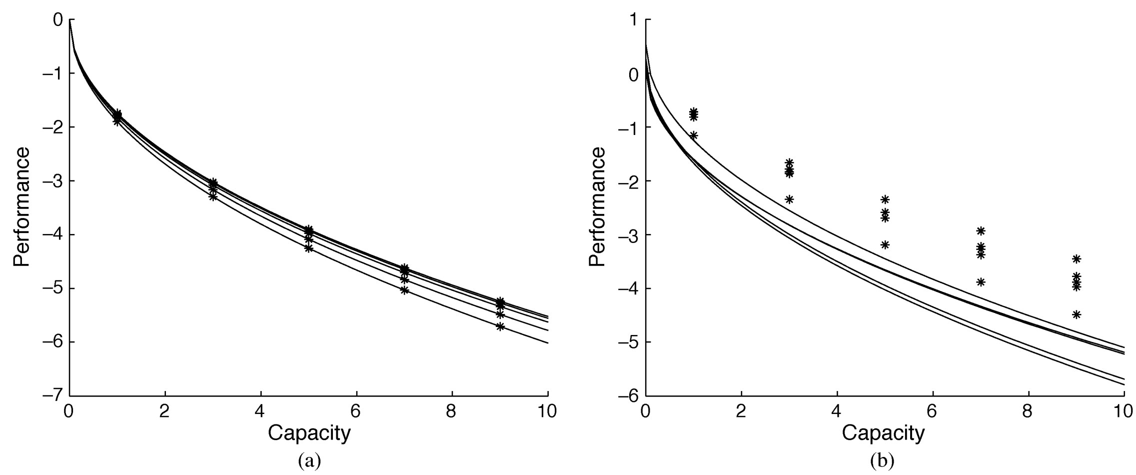

Fig. 5. Analytic bound of Theorem 1 compared to the optimization-based bound of Corollary 2 for random DAGs. The solid curves are the analytic bound performances, and the asterisks are the optimization-based bound performances. Each curve represents a different random graph topology with the number of vertices fixed. (a) $\mu = 0$ case. (b) Random μ case.

We can show that as long $0 \leq \{b, c, d\} \leq 1$:

$$r^2 \leq \max\left\{1 - 2a, \max_{2 \leq i \leq |V|-1} \{i - 2(b+c) - (i-2)d\}\right\} + \left(a^2 + (|V|-2)(b^2+c^2) + \frac{(|V|-2)(|V|-3)}{2} d^2\right).$$

Note that $((|V|-1) - 2(b+c) - (|V|-3)d)$ is the length (using $(1-2c)$ as the edge weight vector) of the acyclic path having the most edges.

If we set $a = 1/2$ and $0 \leq \{b, c, d\} \leq 1$, we get

$$r^2 \leq ((|V|-1) - 2(b+c) - (|V|-3)\,d) + \left(\frac{1}{4} + (|V|-2)(b^2+c^2) + \frac{(|V|-2)(|V|-3)}{2} d^2\right).$$

Maximizing over b, c, and d, we get $b = c = d = 1/|V| - 2$, and

$$(r_o^*)^2 \leq \frac{4|V|^2 - 13|V| + 4}{4(|V|-2)} \approx |V| \approx \sqrt{|E|}$$

where the final approximate equality is specific to the complete graph. Therefore, the average performance is lower bounded by $J(C) \gtrsim J(0) - \sqrt[4]{|E|}\sqrt{C}$.

Finally, we present two examples comparing Theorem 1 to Corollary 2. In both, we consider two cases for the mean μ: 1) $\mu = 0$ and 2) a random μ. Because our analytic bounds grow unbounded for increasing C (even for $C > \mathrm{VAR}[W]$), we compare the bounds only over $0 \leq C \leq \min_e\{\mathrm{VAR}[W_e]\}$. This will allow us to examine the impact of increasing capacity without saturating information from any one edge.

Example 7: In this example, we consider a graph having two disjoint paths from s to t and consisting of the same number of edges, and where $\mathrm{VAR}[W_e] \geq 10$ for all edges. We compute performance bounds over $0 \leq C \leq 10$. Fig. 4(a) and (b) shows the performances of two approaches in each case. For the case $\mu = 0$, the performances are identical over capacity and graph size. In the case of a random (nonzero) μ, the bound of Corollary 2 has slightly better performance, but it has roughly the same rate of performance improvement.

Example 8: In this example, we consider random DAGs consisting of ten vertices and where $\mathrm{VAR}[W_e] \geq 10$ for all edges. We compute performance bounds over $0 \leq C \leq 10$.

Fig. 5(a) and (b) shows the performances of two approaches in each case. For the case $\mu = 0$, the performances are identical over capacity and graph topology. In the case of a random (nonzero) μ, the bound of Corollary 2 has slightly better performance, but, once again, has roughly the same rate of performance improvement.

Remark 8: There are two interesting observations in the computational examples. The first is that the bounds from both Theorem 1 and Corollary 2 seem to be identical in the regime $C \leq \min_e\{\mathrm{VAR}[W_e]\}$ and $\mu = 0$, despite applying different outer-approximations in their respective formulations (Theorem 1 outer approximates $\mathcal{P}$ while Corollary 2 outer approximates $\Gamma(C)$). The second observation is that in the case of nonzero μ, the bounds at least seem to possess the same rate of improvement with increasing capacity.

VII. Conclusion and Future Work

A general framework for studying stochastic optimizations under partial information was presented and specialized to shortest path optimization. As part of this specialization, a convenient parametrization for information capacity was used that allowed us to conveniently optimize information. A new graph characterization for shortest path optimization was used to provide new algorithms (convex optimizations in the case of Gaussian edge weights) for information optimization as well as provide analytic performance bounds.

Future work includes examining other useful approximate geometries for $\mathcal{P}$ and examining the relationship between such geometries with the distributions resulting from Corollary 2. We also plan to study a generalization of our framework by which the agent is able to receive information as it traverses the graph. In this setting, we seek to determine how delayed but possibly optimized information impacts performance relative to obtaining all information upfront. Finally, methods for determining the set of (in)actionable information would be useful in general. The set of inactionable information in particular can be explicitly characterized as the set $\mathcal{W}$ for the perturbation W such that $h(x, w) = h(x', w)$ for all $x, x' \in \mathcal{X}$ and for each $w \in \mathcal{W}$.

Appendix

II. Proofs of Results

Proof of Theorem 4: We prove the result by providing such an algorithm. First, define

$$f\left(\{n^1, \ldots, n^i\}, j\right) = \underset{n \in \Re^{|E|}, \overline{J}: V \to \Re, \underline{J}: V \to \Re}{\arg\min} \|n\| \text{ subject to}$$

$$n_j = 1, \left(n^k\right)^T n = 0 \text{ for all } k \leq i$$

$$\overline{J}(v) = \max_{e|\mathrm{hd}(e)=v}\left\{\overline{J}\left(\mathrm{tl}(e)\right) + n_e\right\}$$

$$\underline{J}(v) = \min_{e|\mathrm{hd}(e)=v}\left\{\underline{J}\left(\mathrm{tl}(e)\right) + n_e\right\}$$

$$\overline{J}(s) = \underline{J}(s), \overline{J}(t) = \underline{J}(t) = 0$$

with the definition $f(\{n^1, \ldots, n^{i-1}\}, j) = 0$ if the optimization is infeasible.

The algorithm is as follows:

$H_{\mathcal{P}} \leftarrow I$; $i \leftarrow 1$; $j \leftarrow 1$.
while $j \leq |E|$ **do**.
  $n^* \leftarrow f(\{n^1, \ldots, n^{i-1}\}, j)$
  if $n^* \neq 0$ **then**.
    $H_{\mathcal{P}} \leftarrow H_{\mathcal{P}} - (n^*(n^*)^T / \|n^*\|^2)$.
    $n^i \leftarrow n^*$.
    $i \leftarrow i + 1$.
  else.
    $j \leftarrow j + 1$.
  end if.
end while.

It is clear that the above algorithm terminates in a polynomial number of steps, so we only need to prove that it generates a valid projection matrix $H_{\mathcal{P}}$. For any orthogonal basis $\{n^1, n^2, \ldots, n^m\}$ of $S_{\mathcal{P}}^{\perp}$, a projection matrix for $S_{\mathcal{P}}$ can be written as $I - \sum_{i=1}^m (n^i (n^i)^T / \|n^i\|^2)$. Therefore, we need to show that the set $\{n^i\}$ generated from the algorithm is such a basis.

A vector $n \in \Re^{|E|}$ being normal to $\mathcal{S}_{\mathcal{P}}$ is equivalent to $n^T(\mathcal{P} - \overline{p}) = \{0\} \Leftrightarrow n^T(p - \overline{p}) = 0$ for all $p \in P$. Therefore, a sufficient and necessary condition for n to be normal is that all paths from s to t have the same length when the edge weights are n.

The longest path in the graph when the edge weights are given by n is the unique function $\overline{J} : V \to \Re$ satisfying $\overline{J}(v) = \max_{e|\mathrm{hd}(e)=v}\{\overline{J}(\mathrm{tl}(e)) + n_e\}$. The shortest path is the unique function $\underline{J}$ satisfying the same equality with $\max \to \min$. An equivalent condition to all paths having the same length is $\overline{J}(s) = \underline{J}(s)$. Therefore, if f is feasible, any optimal solution vector n^* provided by f must be normal to $\mathcal{S}_{\mathcal{P}}$, and, further, it must be nonzero. Therefore, any set of such feasible vectors $\{n^1, n^2, \ldots, n^k\}$ must lie in $S_{\mathcal{P}}^{\perp}$.

By the orthogonality constraint $(n^k)^T n = 0$ and the constraint $n_j = 1$, any set of vectors $\{n^1, n^2, \ldots, n^k\}$ yielded by the algorithm must be nonzero and orthogonal; hence, the set is a subset of an orthogonal basis for $S_{\mathcal{P}}^{\perp}$.

Now, suppose the set $\{n^1, n^2, \ldots, n^k\}$ generated by the algorithm is a strict subset of a basis. Then there is a nonzero vector n that is orthogonal to each n^j and lies in $S_{\mathcal{P}}^{\perp}$. Let i be the smallest index of the nonzero components of n, and assume without loss of generality that $n_i = 1$ (n can be normalized to produce this). Finally, let $l = \arg\max_j\{n^j | n_i^j \neq 0\}$.

By the definition of l, $f(\{n^1, \ldots, n^l\}, i)$ is infeasible since, otherwise, $l = l + 1$, a contradiction. By the existence of n, though, we know that $f(\{n^1, \ldots, n^l\}, i)$ is feasible since n satisfies all of the constraints of the optimization. The contradiction implies that $\{n^1, n^2, \ldots, n^k\}$ must be an orthogonal basis for $S_{\mathcal{P}}^{\perp}$. Hence, $H_{\mathcal{P}}$ is a projection matrix for $S_{\mathcal{P}}$. ■

Proof of Corollary 2: We present a detailed sketch of the proof. First, if we remove the capacity constraint from (8) and instead fix the variances $\mathrm{VAR}\hat{W}_e = \lambda_e^2$, we get

$$\underline{J}(\{\lambda_e\}) = \min_{p_{\hat{W}}}\left\{\mathrm{E}\left[\min_p\{p^T \hat{W}\}\right]\right\}$$

$$\text{subject to } \mathrm{E}[\hat{W}] = \mu, \mathrm{VAR}\,[\hat{W}] = \lambda_e^2.$$

For ease, we denote the lower bound for performance in this case as $\underline{J}(\lambda)$ for the vector $\lambda = (\lambda_e)_e$. This optimization is of the form Equation (3.7) in [10]. By Theorem 3.1 in [10], it is equivalent to Equation (3.8) in [10]. Substituting our constraints yields a quadratic objective and second-degree polynomial inequalities, which we can re-express as operations and inequalities on semi-definite matrices:

$$\underline{J}(\lambda) = \max_{\{G_e\}, d \in \Re^{|E|}} \left\{ J(d) + \sum_e G_e \cdot \begin{bmatrix} \lambda_e^2 + \mu_e^2 & \mu_e \\ \mu_e & 1 \end{bmatrix} \right\}$$
$$\text{subject to } G_e \leq \left\{ 0, \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & -d_e \end{bmatrix} \right\}$$

where by our definition of $J(W)$, $J(d)$ is simply $J(W)$ with $W = d$.

Let $\Gamma'(C) = \{\lambda \in \Re^{|E|} | 0 \leq \lambda_e^2 \leq \sigma_e^2, \|\lambda\|_2^2 \leq C\}$. $\Gamma'(C)$ is clearly a convex set. It is also clear that

$$\underline{J}(C) = \min_{\lambda \in \Gamma'(C)} \underline{J}(\lambda).$$

Taking the dual of the optimization for $\underline{J}(\gamma)$ with respect to $\{G_e\}$ yields a new inner optimization

$$\underline{J}(\lambda) = \min_{H_e} \max_d \left\{ J(d) - \sum_e H_e \cdot \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & -d_e \end{bmatrix} \right\} \text{ subject to}$$
$$H_e \leq 0, H_e \geq - \begin{bmatrix} \lambda_e^2 + \mu_e^2 & \mu_e \\ \mu_e & 1 \end{bmatrix}.$$

Let $H_e = \begin{bmatrix} a_e & b_e \\ b_e & c_e \end{bmatrix}$. Letting $c = [c_1 \ldots c_{|E|}]^T$ and $d = [d_1 \ldots d_{|E|}]^T$, the objective in the minimax is

$$\begin{aligned} J(d) + c^T d - \sum_e \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \cdot H_e \\ = \min_{p \in \mathcal{P}} \{p^T d\} + c^T d - \sum_e \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \cdot H_e \\ = \min_{p \in \mathcal{P}} \left\{ (p+c)^T d \right\} + \sum_e \begin{bmatrix} 0 & -\frac{1}{2} \\ -\frac{1}{2} & 0 \end{bmatrix} \cdot H_e \end{aligned}$$

Of interest to us is the minimaximin expression:

$$\min_c \max_d \min_{p \in \mathcal{P}} \left\{ (p+c)^T d \right\}$$

If $-c \in \mathcal{P}$, then this expression must always be nonpositive since $0 \in \mathcal{P} - c$. If $-c \notin \mathcal{P}$, then one can show that the expression will always be ∞. Therefore, we require $-c \in \mathcal{P}$. In this case, $d = 0$ is the optimal strategy for d since that maximizes the expression to 0.

The constraint $c \in \mathcal{P}$ is represented by

$$\sum_e \left(\begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \cdot H_e \right) v_e \in \mathcal{P}.$$

The remainder of the claim follows. ∎

REFERENCES

[1] C. C. Huang, I. Vertinsky, and W. T. Ziemba, "Sharp bounds on the value of perfect information," *Oper. Res.*, vol. 25, no. 1, pp. 128–139, 1977.

[2] M. Avriel and A. C. Williams, "The value of information and stochastic programming," *Oper. Res.*, vol. 18, no. 5, pp. 947–954, 1970.

[3] V. Firchau, "Bounds for the expected value of information," *OR Spectrum*, vol. 2, no. 2, pp. 65–73, Jun. 1980.

[4] A. Ben-Tal and E. Hochman, "Stochastic programs with incomplete information," *Oper. Res.*, vol. 24, no. 2, pp. 336–347, 1976.

[5] S. Walley, H. Tan, and A. Viterbi, "Shortest path cost distribution in random graphs with positive integer edge costs," in *Proc. 12th Annu. Joint Conf. IEEE Comput. Commun. Soc.*, Mar. 1993, pp. 1023–1032.

[6] R. Hassin and E. Zemel, "On shortest paths in graphs with random weights," *Math. Oper. Res.*, vol. 10, no. 4, pp. 557–564, 1985.

[7] G. Weiss, "Stochastic bounds on distributions of optimal value functions with applications to pert, network flows and reliability," *Oper. Res.*, vol. 34, no. 4, pp. 595–605, 1986.

[8] I. Meilijson and A. Nadas, "Convex majorization with an application to the length of critical paths.," *J. Appl. Probab.*, vol. 16, pp. 671–677, 1979.

[9] L. P. Devroye, "Inequalities for the completion times of stochastic pert networks," *Math. Oper. Res.*, vol. 4, no. 4, pp. 441–447, 1979.

[10] D. Bertsimas, K. Natarajan, and C.-P. Teo, "Probabilistic combinatorial optimization: Moments, semidefinite programming, and asymptotic bounds," *SIAM J. Optim.*, vol. 15, no. 1, pp. 185–209, 2005.

[11] K. Isii, "On sharpness of tchebycheff-type inequalities," *Ann. Inst. Statist. Math.*, vol. 14, no. 1, pp. 185–197, 1962.

[12] R. Cogill, "Randomized load balancing with non-uniform task lengths," in *Proc. 45th Annu. Allerton Conf.*, Sep. 2007.

[13] J. H. Hagstrom, "Computational complexity of pert problems," *Networks*, vol. 18, pp. 47–56, 1988.

[14] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[15] D. Bertsimas, K. Natarajan, and C.-P. Teo, "Tight bounds on expected order statistics," *Probab. Eng. Inf. Sci.*, vol. 20, no. 4, pp. 667–686, 2006.

[16] M. Rinehart and M. A. Dahleh, "Shortest path optimization under limited information," in *Proc. 48th IEEE Conf. Decision Control*, Dec. 2009, pp. 2064–2069.

[17] K. Natarajan, C. Teo, and Z. Zheng, "Mixed zero one linear programs under objective uncertainty: A completely positive representation," *Oper. Res.*, to be published.

Michael Rinehart received the B.S. degree in electrical engineering from the University of Maryland, College Park, in 2003 and the S.M. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, in 2005 and 2009, respectively.

He is currently a Senior Research Engineer at BAE Systems Advanced Information Technologies as well as a Research Affiliate with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge.

Munther A. Dahleh was born in 1962. He received the B.S. degree from Texas A&M University, College Station, in 1983 and the Ph.D. degree from Rice University, Houston, TX, in 1987, all in electrical engineering.

Since the Ph.D. degree, he has been with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge. He is coauthor of the books *Control of Uncertain Systems: A Linear Programming Approach* (Prentice-Hall, 1995) and *Computational Methods for Controller Design* (Springer, 1998). His research interests include the interface of robust control, information theory, and computation; model reduction of discrete-alphabet hidden Markov models; universal learning approaches for systems with continuous and discrete alphabets; and the interface between systems theory and neurobiology.

Dr. Dahleh was the recipient George Axelby Outstanding Paper Award (1987, 2004, and 2010), an NSF Presidential Young Investigator Award (1991), the American Control Council's Donald P. Eckman Award (1993), the Graduate Student Council Teaching Award (1995), and the Hugo Schuck Award for Theory (2006).