Joint 48th IEEE Conference on Decision and Control and
28th Chinese Control Conference
Shanghai, P.R. China, December 16-18, 2009

Shortest Path Optimization under Limited Information

Michael Rinehart and Munther A. Dahleh

***Abstract*— The problem of finding an optimal path in an uncertain graph arises in numerous applications, including network routing, path-planning for vehicles, and the control of finite-state systems. While techniques in robust and stochastic programming can be employed to compute, respectively, worst-case and average-optimal solutions to the shortest-path problem, we consider an alternative problem where the agent that traverses the graph can request limited information about the graph before choosing a path to traverse.**

In this paper, we define and quantify a notion of information that is compatible to this performance-based framework, bound the performance of the agent subject to a bound on the capacity of the information it can request, and present algorithms for optimizing information.

I. Introduction

Path planning on uncertain graphs is considered in numerous areas of study, and specific studies of the problem have appeared many times in the literature. For example, determining the complexity of finding the shortest path, computing the distributions of path lengths for certain graph structures or edge-weight distributions [1][2], and bounding the mean of the shortest path [3][4]. In this paper, we consider an alternate formulation where the graph's edge weights are random, but the agent that traverses the graph is limited to knowing them within a specified degree of accuracy. This formulation is a special case of uncertain decision problems whereby a random objective is to be minimized subject to some prior information about the realization of the objective.

We begin this paper with a brief presentation of stochastic optimization but under a particular definition of partial information. We then immediately specialize this framework to our problem of interest: shortest-path optimization under limited information. For this application, we will provide performance bounds for the shortest path algorithm as well as present algorithms for (sub)optimizing partial information. We conclude the paper with some final remarks about future work.

II. Definitions, Notation, and Formulation

A. Random Variables and Sets

We write random variables (RVs) in capital letters (e.g., X), and denote the event $X \in S$ as $P(X \in S)$ or $P_X(S)$. We write $X \sim p$ if X has p as its probability density function (PDF), and 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 $X = (X_1, \ldots, X_n)$, $\mathrm{VAR}[X] = \sum_{i=1}^{n} \mathrm{VAR}[X_i]$ and $\mathrm{COV}[X] = \mathrm{E}\left[XX^T\right] - \mathrm{E}[X]\,\mathrm{E}[X]^T$. 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 $X \overset{\mathrm{d}}{=} Y$ if both RVs are drawn from the same distribution.

For two sets A and B, $A - B$ is the set of elements in A but not in B. Finally, $|A|$ is the number of elements in A.

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 $w = \left[w_1 \ldots w_{|E|}\right]^T$. Because we consider edge weights to be random, we write the vector as W, and we assume that the probability distribution is known. Finally, we denote the first and second moments of W by

$$\mu = \mathrm{E}[W]\,,\ \mu_e = \mathrm{E}[W_e]\,,$$
$$\Lambda = \mathrm{COV}[W]\,,\ \sigma_e^2 = \Lambda_{ee} = \mathrm{VAR}[W_e]\,.$$

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

Definition 1 (Path): A sequence $p = (e_1, e_2, \ldots, e_n)$ 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 $p = (e_i)$ is acyclic if there are no two indices $i < j$ such that $\mathrm{hd}(e_i) = \mathrm{tl}(e_j)$.

Assumption 1 (DAG): G is a directed acyclic graph (DAG) (i.e., all paths the p of G are acyclic).

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

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

Let $P = P(G)$ be the set of all paths from s to t in G. 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 a set of all such vectors in $\Re^{|E|}$. Let $\mathcal{P} = $ convex hull$\{P\}$. It is well known that $\mathcal{P}$ has an efficient representation as a finite number of equality

Michael Rinehart (mdrine@mit.edu) and Munther A. Dahleh (dahleh@mit.edu) are with the Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

978-1-4244-3872-3/09/$25.00 ©2009 IEEE

constraints:

$\mathcal{P} = x \in [0,1]^{|E|}$ such that

$$\sum_{e|e=\mathrm{tl}(v)} x_e - \sum_{e|e=\mathrm{hd}(v)} x_e = \begin{cases} 1 & v = s \\ -1 & v = t \\ 0 & \text{otherwise.} \end{cases}$$

Using our vector notation, the length of a path $p \in P$ is simply $p^T W$.

C. Stochastic Optimization under Partial Information

To motivate our framework for shortest path optimization under limited information, we present the basic ideas in a general stochastic optimization setting. For the purposes of this section, let W be any RV with some 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 $\mathrm{E}\left[J(W)\right]$.

Consider now the task of finding an "optimal" decision x without knowing W. Of course, 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, $\mathrm{E}\left[\overline{J}\right] = \overline{J}$. By Jensen's Inequality, $\mathrm{E}\left[J(W)\right] \leq \overline{J}$.

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 our *information* about W and write the optimization as a function of our information:

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

Once again, because Y is a RV, $J(Y)$ is also a RV, and so the average performance under Y is simply $\mathrm{E}\left[J(Y)\right]$. Clearly, the information Y contains about W is completely determined by their joint-distribution p_{WY}, so we define $J(p_{WY}) = \mathrm{E}\left[J(Y)\right]$:

$$J(p_{WY}) = \mathrm{E}\left[\min_{x \in \mathcal{X}} \mathrm{E}\left[h(x, W)|Y\right]\right]. \tag{1}$$

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 $J(Y) = J(W)$.

A goal of this paper is to choose the information Y so as to minimize the average performance $J(p_{WY})$ under some constraint on the "total information" Y has about W. We can bound information simply by placing a constraint on the set of allowable joint-distributions p_{WY}. Formally, if we let Γ be such a constraint set for p_{WY}, we seek the solution to

$$J(\Gamma) = \min_{p_{WY} \in \Gamma} J(p_{WY}). \tag{2}$$

We call $J(\Gamma)$ the *optimal performance under* Γ, and we call (2) the *information optimization*.

We can further generalize the concept of information optimization to optimizations over a family of constraint sets $\{\Gamma(C)\}$. Here, the scalar C is called the *capacity*, and (for no technical reason) we assume that $\Gamma(C_1) \subset \Gamma(C_2)$ whenever $C_1 \leq C_2$. For ease, we simplify our notation by writing $J(C) = J(\Gamma(C))$ and call $J(C)$ the *optimal performance under capacity* C.

Remark 2: The "natural" choice for $\{\Gamma(C)\}$ depends on the application. For example, if we were to specialize (2) to a rate-distortion problem, we could consider minimizing the distortion $d(\hat{W}(Y), W)$ between $\hat{W}(Y)$ and W under the set $\Gamma(C) = \{p_{WY} | I(Y; W) \leq C\}$.

D. Specializing to Shortest Path Optimization

We now specialize our framework (2) to shortest path optimization. First, for any RV Y, (1) specializes to

$$\begin{aligned} J(p_{WY}) &= \mathrm{E}\left[\min_{p \in P} \left\{p^T \mathrm{E}\left[W|Y\right]\right\}\right] \\ &= \mathrm{E}\left[\min_{p \in P} \left\{p^T \hat{W}\right\}\right] \end{aligned} \tag{3}$$

All that is left is to define our information constraint sets. As we will see, a particularly convenient set of capacity constraints is

$$\Gamma(C) = \{p_{WY} \mid \mathrm{VAR}\left[\hat{W}\right] \leq C\}. \tag{4}$$

Notice that both (3) and (4) both only depend on $p_{\hat{W}}$, not p_{WY} [1]. Therefore, from here on, all of our statements will be made with respect to knowing $p_{\hat{W}}$ only.

E. Objective

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

III. Information Optimization

A challenge in information optimization is the evaluation of the objective $J(p_{\hat{W}})$. To avoid this difficulty, we instead seek to optimize upper and lower bounds for the objective. In the case of an optimizing an upper bound, of course, one is guaranteed some improvement in actual performance.

We begin this section with some notation for the case of Gaussian edge weights. We then present two examples that highlight the impact that even a modest information optimization can have. We then present two optimization algorithms and conclude the section with a technique for reducing the complexity of information optimization.

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

A. Simplifying to the Gaussian Case

Consider the case of (a) independent edge weights and (b) W and Y being jointly-Gaussian so that $\hat{W}(Y)$ is Gaussian as well. If we let

$$\gamma_e^2 = \mathrm{VAR}\left[\mathrm{E}\left[W_e|Y\right]\right],$$

then $\hat{W}_e(Y) = \gamma_e Z_e + \mu_e$ for independent RVs $Z_e \sim N(0,1)$. Because μ_e is fixed, the distribution of $\hat{W}$ only varies with γ_e. We now no longer need to consider optimizing the distribution $p_{\hat{W}}$ but rather only over the variances $\{\gamma_e^2\}$. Furthermore, we can write

$$\Gamma(C) = \{\gamma = (\gamma_e)_e \text{ such that } \gamma \geq 0, \|\gamma\|^2 \leq C\},$$

so that $\Gamma(C)$ is a convex set. By parameterizing over γ instead of $p_{\hat{W}}$, we write the average performance as $J(\gamma)$ instead of $J(p_{\hat{W}})$.

B. The Impact of Information Selection: Comparative Examples

Using our notation from the Gaussian case, we can highlight the impact that even a modest information optimization can have.

Example 1: Let G be the graph having n disjoint paths $\{P_1, \ldots, P_n\}$ from s to t with each path having n edges each, and let $W_{ij} \sim N(0,1)$ be the (random) weight of edge j on path i. Let the "distribution" $\gamma^s \in \Gamma(n)$ satisfy $\gamma_e^s = 1$ if $e \in P_1$ and $\gamma_e^s = 0$ 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 $e \notin P_1$.

For this distribution, the average performance is

$$\begin{aligned} J(\gamma^s) &= \mathrm{E}\left[\min\left\{\sum_j W_{1j}, 0\right\}\right] \\ &= \mathrm{E}\left[\min\left\{\sum_j \gamma_{1j} Z_{1j}, 0\right\}\right] \\ &= \mathrm{E}\left[\min\left\{\sqrt{\sum_j \gamma_{1j}^2} Z, 0\right\}\right] \\ &= \mathrm{E}\left[\min\left\{\sqrt{n} Z, 0\right\}\right] \\ &= \sqrt{n}\,\mathrm{E}\left[\min\left\{Z, 0\right\}\right], \end{aligned}$$

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

Example 2: Let G be the same graph as in Example 1, but take a different distribution $\gamma^p \in \Gamma(n)$ such that $\gamma_e^p = 1$ if e is the first edge of any path P_i and $\gamma_e^p = 0$ 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.

For this distribution, the average performance is

$$\begin{aligned} J(\gamma^p) &= \mathrm{E}\left[\min\left\{W_{11}+0, W_{21}+0, \ldots, W_{n1}+0\right\}\right] \\ &= \mathrm{E}\left[\min\left\{Z_{11}, Z_{21}, \ldots, Z_{n1}\right\}\right] \\ &\geq -\sqrt{2\ln n} \end{aligned}$$

where the last inequality is obtained by using Lemma 3 in [5].

There is a significant difference between the performances of the two examples. If we increase n, the average performance yielded from applying γ^s outstrips that obtained using γ^p quite substantially.

C. Information Optimization via Lower Bounds

An approach for "optimizing" information is to optimize a lower bound for $J(p_{\hat{W}})$. One method to obtain a lower bound is to compute the lower bound via an optimization over arbitrary distributions subject only to a few constraints. For this purpose, we can apply an extension of the generalized Chebyshev inequality presented in [6].

For our problem, it is convenient (and computationally efficient) to compute such a bound directly for $J(C)$ subject to 1st order (the mean) and 2nd order (the variance and capacity bound) constraints on the distribution:

$$\begin{aligned} \underline{J}(C) = \min_{p_{\hat{W}}} &\left\{\mathrm{E}\left[\min_p\left\{p^T \hat{W}\right\}\right]\right\} \text{ subject to} \\ &\mathrm{E}\left[\hat{W}\right] = \mu,\ \mathrm{VAR}\left[\hat{W}\right] \leq C \\ &0 \leq \mathrm{VAR}\left[\hat{W}_e\right] \leq \sigma_e^2. \end{aligned} \tag{5}$$

Clearly, $J(C) \geq \underline{J}(C)$. It can be shown that the dual to this optimization is a semi-definite optimization having a number of constraints larger than the number of paths in the graph (each path generates a constraint). Hence, it is impractical for even moderately-sized graphs.

An approach for practically bounding the performance of uncertain 0-1 optimizations is presented in [7], but, in our case, it requires some modification. This is because [7] assumes that feasible distributions for $p_{\hat{W}}$ are constrained in their marginals only, but our capacity constraint in (5) disobeys this rule. Nonetheless, the approach is still useful, and we present our first information optimization algorithm as a corollary to Corollary 3.2 in [7].

Corollary 1:

$$\begin{aligned} \underline{J}(C) = \min &\left\{\sum_e \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \cdot H_e\right\} \text{ subject to} \\ &H_e \geq 0 \\ &H_e \geq \begin{bmatrix} \gamma_e^2 + \mu_e^2 & \mu_e \\ \mu_e & 1 \end{bmatrix} \\ &\sum_e \left(\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \cdot H_e\right) v_e \in \mathcal{P} \\ &0 \leq \gamma_e^2 \leq \sigma_e^2 \\ &\sum_e \gamma_e^2 \leq C \end{aligned} \tag{6}$$

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

Proof: For space reasons, we present only a detailed sketch of the proof. First, if we remove the capacity constraint from (5) and instead fix the variances $\{\gamma_e^2\}$ for $\hat{W}$, we get a tight lower bound for $J(p_{\hat{W}})$ for any $p_{\hat{W}}$ having variances given by the γ_e's. By Corollary 3.2 in [7], this lower bound is equivalent to [2]

$$\underline{J}(\gamma) = \max_{\{\Gamma_e\}, d \in \Re^{|E|}} \left\{ J(d) + \sum_e \Gamma_e \cdot \begin{bmatrix} \gamma_e^2 + \mu_e^2 & \mu_e \\ \mu_e & 1 \end{bmatrix} \right\}$$

subject to

$$\Gamma_e \leq \{0, \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & d_e \end{bmatrix}\}.$$

Now we clearly have

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

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

$$\min_{H_e} \max_{d} \left\{ J(d) + \sum_e H_e \cdot \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & -d_e \end{bmatrix} \right\} \quad \text{subject to}$$

$$H_e \geq 0$$

$$H_e \geq \begin{bmatrix} \gamma_e^2 + \mu_e^2 & \mu_e \\ \mu_e & 1 \end{bmatrix}.$$

Let $H_e = \begin{bmatrix} a_e & b_e \\ b_e & c_e \end{bmatrix}$. Then the objective in the minimax is

$$\begin{aligned} &= J(d) + \sum_e b_e - \sum_e c_e d_e \\ &= \min_{p \in \mathcal{P}} \{p^T d\} - c^T d + \sum_e \begin{bmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \cdot H_e \\ &= \min_{p \in \mathcal{P}} \{(p-c)^T d\} + \text{constant}. \end{aligned}$$

Of interest to us is the minimaximin expression:

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

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 $c \in \mathcal{P}$. Since $\mathcal{P}$ has no volume (not proven in this paper), d should make itself orthogonol to the subspace containing $\mathcal{P} - c$ (which is the same for all c) in order to maximize 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 easily follows. ■

Remark 3: By our definition of $J(W)$, $J(d)$ is simply $J(W)$ with $W = d$, a constant.

A significant drawback to this lower bound optimization is that the resulting lower bound can be quite conservative. In particular, the optimization over distributions for $p_{\hat{W}}$ may include non-realizable distributions. For example, in the case of parallel requests, we know that the rate of improvement with each request is slower than logarithmic for the Gaussian case, but the minimizing distribution yields a square root improvement [8] since the optimization generates dependencies among the edge weights. Hence, it is not clear if this information optimization is meaningful, despite it giving a potentially useful lower bound for the performance.

[2] Although we are not applying any Gaussian assumptions, because we are restricted to first and second order moments, we adopt the same approach of writing $J(\gamma)$ since the optimization is only a function of this characteristic of the distribution.

D. Information Optimization via Upper Bounds

We now consider an alternative information optimization that optimizes an upper bound for $J(p_{\hat{W}})$ that better leverages the structure of the graph to reduce complexity. The upper bound is based in Jensen's Inequality, and it applies a dynamic-programming-like approach to computing the average performance backward through the graph, but it requires that the edge weights obey the Gaussian assumptions discussed before.

Theorem 1: An upper bound $J(C)$ under independent Gaussian edge weights is

$$J(C) \leq \min_{\{\gamma_e^2\} \in \Gamma(C)} \{\overline{J}(s, \{\gamma_e^2\})\} \tag{7}$$

where

$$\overline{J}(v, \{\gamma_e^2\}) = \int j \frac{\partial}{\partial x} \cdot$$

$$\left(\prod_{e|\,\mathrm{hd}(e)=v} \left[1 - \Phi\left(\frac{x - \overline{J}(\mathrm{tl}\,(e)\,, \{\gamma_e^2\}) - \mu_e}{\gamma_e} \right) \right] \right) |_{x=j} \mathrm{dj}$$

and $\overline{J}(t, \{\gamma_e^2\}) = 0$.

Proof: Let $J(v, \hat{W})$ be the length of the shortest path from vertex v to t under edge weights $\hat{W}$:

$$J(v, \hat{W}) = \min_{e|\,\mathrm{hd}(e)=v} \left\{ J(\mathrm{tl}\,(e)\,, \hat{W}) + \hat{W}_e \right\},$$

$$J(t, \hat{W}) = 0.$$

By Jensen's Inequality and acyclicity, we have

$$\mathrm{E}\left[J(v, \hat{W})\right] \leq \mathrm{E}\left[\min_{e|\,\mathrm{hd}(e)=v} \left\{ \mathrm{E}\left[J(\mathrm{tl}\,(e)\,, \hat{W})\right] + \hat{W}_e \right\} \right],$$

so that the set of equations

$$\overline{J}(v, \{\gamma_e^2\}) = \mathrm{E}\left[\min_{e|\,\mathrm{hd}(e)=v} \left\{ \overline{J}(\mathrm{tl}\,(e)\,, \{\gamma_e^2\}) + \hat{W}_e \right\} \right]$$

yields an upper bound for $\mathrm{E}\left[J(v, \hat{W})\right]$.

Now let,

$$\overline{J}(v, \{\gamma_e^2\}, \hat{W}) = \min_{e|\,\mathrm{hd}(e)=v} \left\{ \overline{J}(\mathrm{tl}\,(e)\,, \{\gamma_e^2\}) + \hat{W}_e \right\}$$

so that $\overline{J}(v, \{\gamma_e^2\}) = \mathrm{E}\left[\overline{J}(v, \{\gamma_e^2\}, \hat{W})\right]$.

Using

$$P(\hat{W}_e > x) = 1 - \Phi\left(\frac{x - \mu_e}{\gamma_e}\right)$$

and $P(\min_i \{X_i\} > x) = \prod_i P(X_i > x)$ for independent RVs $\{X_i\}$ allows us to compute the CDF for the

$\overline{J}(v, \{\gamma_e^2\}, \hat{W})$. Taking a derivative and integrating yields its expected value:

$$\overline{J}(v, \{\gamma_e^2\}) = \int j \frac{\partial}{\partial x} \cdot$$
$$\left(\prod_{e|\,\mathrm{hd}(e)=v} \left[1 - \Phi\left(\frac{x - \overline{J}(\mathrm{tl}\,(e)\,, \{\gamma_e^2\}) - \mu_e}{\gamma_e} \right) \right] \right) |_{x=j} \mathrm{dj}$$

■

E. Reducing Complexity by Path Pruning

The information optimizations presented above are quite manageable, but we can further improve their performances by leveraging a simple fact about real-world graphs: paths of long average length are almost never the shortest path, so information about them can be neglected. In this section, we provide a bound on the performance lost from pruning such paths.

Define $\Theta_X(k) = \mathrm{E}[\min\{X, k\}]$. If $X \sim N(0,1)$, then $\Theta_X(0) = \frac{-1}{\sqrt{2\pi}}$. We use Θ to write a very simple pruning algorithm. Let p^* be the path in the graph with shortest average length. We want to prune the paths p that are often not the shortest path when compared to p^*. Formally, we prune every path p satisfying

$$\mathrm{E}\left[\min\left\{(p^*)^T W, p^T W\right\}\right] - (p^*)^T \mu = \Theta_{(p-p^*)^T W}(0) \approx 0.$$

However, it may be the case that no path in the graph is pruned regardless of length because the variances of such paths can be very large (a path with a sufficiently large variance has a good chance of being the shortest path even if the mean length is long). As a fix, we can bound the variance of a path as a function of its mean length. A particularly useful bound is $\mathrm{VAR}\left[p^T W\right] \le \mathrm{E}\left[p^T W\right]$, a sufficient condition for which is $\sigma_e^2 \le \mu_e$ and which is easy to check.

Under this assumption, we can prune paths simply by looking at their mean lengths, yielding the follow low-complexity pruning algorithm:

1) Compute the average shortest path length from each vertex v to t.
2) Remove those vertices whose lengths exceed some predetermined length L.

We now provide a bound on the performance lost from path pruning. For space reasons, we do not provide a proof of this result and we restrict it to the Gaussian case.

Theorem 2: Assume $\mathrm{VAR}\left[p^T W\right] \le \mathrm{E}\left[p^T W\right]$ and that the edge weights are independent Gaussian. The performance lost from pruning paths p satisfying $\mathrm{E}\left[p^T W\right] \ge L$ is

$$\sum_{k>L} \left[(\text{\# paths of length } k) \left((\overline{L} - k) - \sqrt{k}\,\Theta_Z\left(\frac{\overline{L} - k}{\sqrt{k}} \right) \right) \right]$$

where

$$\lim_{k\to\infty} \left[(\overline{L} - k) - \sqrt{k}\,\Theta_Z\left(\frac{\overline{L} - k}{\sqrt{k}} \right) \right] = 0.$$

A consequence of Theorem 2 is that if the number of paths of length k does not increase too quickly, then the performance loss can be bounded.

IV. An Analytic Relationship between Capacity and Performance

Another objective of our study is to establish an intuitive relationship between information capacity and performance. The previous information optimization algorithms are computational in nature and, consequently, are not quite amendable for this purpose, so we take a different approach. The following proposition is a first step in this direction.

Proposition 1: For any function $f : \mathcal{X} \to \Re$ and any $\overline{\mathcal{X}} \supset \mathcal{X}$,

$$\min_{x\in\mathcal{X}} \{h(x, W)\} \ge \min_{x\in\mathcal{X}} \{f(x)\} + \min_{x\in\overline{\mathcal{X}}} \{h(x, W) - f(x)\}.$$

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

Remark 4: Proposition 1 provides a generalization of the approach taken in [8] to generate a low-complexity optimization for bounding the mean of the minimum order statistic. Let $\mathcal{X}$ be the simplex in $\Re^n$ and let $W = (W_1, \ldots W_n)$ be a random vector in $\Re^n$. The the minimum order statistic is given by $\min_i \{W_i\} = \min_{x\in\mathcal{X}} \left\{x^T W\right\}$. The bound in [8] follows by setting $f(x) = x^T z$ for some vector $z \in \Re^n$ and setting $\overline{\mathcal{X}}$ to the unit cube.

While Proposition 1 does not shed much insight into an analytic relationship between information and performance, clever choices for $f(x)$ and $\overline{\mathcal{X}}$ can be used to provide one.

We now use Proposition 1 to provide an analytic bound for performance.

Theorem 3:

$$J(0) \ge J(C) \ge J(0) - \frac{1}{2}\sqrt{|E|}\sqrt{C}. \tag{8}$$

Proof: Since $\mathcal{P}$ is the convex hull of 0-1 vectors, it is contained in the unit cube. Applying Proposition 1 with $z = \mathrm{E}[W]$ and $\overline{\mathcal{X}}$=the unit cube yields

$$J(C) \ge J(0) - \mathrm{E}\left[\min_{x\in\overline{\mathcal{X}}} \left\{ x^T(\hat{W} - \mu) \right\} \right],$$

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

Without loss of generality, assume $\mu = 0$. Then

$$\begin{aligned}
\mathrm{E}\left[\min_{x\in\overline{\mathcal{X}}} \left\{x^T \hat{W}\right\}\right] &= \mathrm{E}\left[\sum_e \min_{x_e\in[0,1]} \left\{x_e \hat{W}_e\right\}\right] \\
&= \sum_e \mathrm{E}\left[\min\left\{0, \hat{W}_e\right\}\right] \\
&= \sum_e \gamma_e\, \mathrm{E}\left[\min\left\{0, \frac{\hat{W}_e}{\gamma_e}\right\}\right] \\
&\ge -\frac{1}{2}\sum_e \gamma_e \\
&\ge -\frac{1}{2}\sqrt{N_e}\sqrt{\sum_e \gamma_e^2} \\
&= -\frac{1}{2}\sqrt{N_e}\sqrt{C}
\end{aligned}$$

where the first inequality results from Chebyshev's Inequality [8], and the last inequality results from Jensen's Inequality. ■

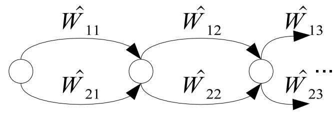

Fig. 1. Graph topology that achieves the lower bound.

The performance bound in Theorem 8 is straight-forward, though somewhat conservative. Because $\mathcal{P}$ for any graph is contained in the unit cube, the lower bound must hold for all possible graphs, so we use no graph topological information in the bound. Interestingly, however, there is a graph topology that does provide the same asymptotic performance as this bound even under independent edge weights. It is presented below.

Theorem 4: There exists a graph with i.i.d. Gaussian edge weights satisfying

$$J(C) = J(0) - \frac{1}{2\sqrt{\pi}}\sqrt{N_e}\sqrt{C}.$$

Proof: Consider the graph in Figure 1. An expression for $J(\gamma)$ is

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

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

Optimizing over the γ_j's (there are $N_e/2$ of them) requires us to solve the optimization

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

for which $\gamma_j = \sqrt{\frac{C}{N_e/2}}$ is the optimum. Thus,

$$J(C) = J(0) + \Theta_Z(0)\frac{N_e}{2}\sqrt{\frac{C}{N_e/2}}.$$

■

In some sense, it should be little surprise that graph topology used to prove Theorem 4 is tight since the number of extreme points of the unit cube and the polytope $\mathcal{P}$ for this graph are on the same order ($\mathcal{P}$ has half as many).

A useful analytic upper bound for performance that uses little distribution information is not known, but we show below that such a bound is trivial when only first and second order moment information about $p_{\hat{W}}$ is used.

Proposition 2: A tight upper bound for $J(C)$ over all distributions subject to a mean and variance constraint is $J(0)$ (the zero-information case).

Proof: Without loss of generality, suppose $\mathrm{E}\left[\hat{W}\right] = 0$ and $\mathrm{VAR}\left[\hat{W}\right] = 1$, and let

$$p_e^n(x) = \frac{1}{2n^2}\delta(x+n) + \frac{1}{2n^2}\delta(x-n) + \frac{2n^2-2}{2n^2}\delta(x).$$

Each distribution p_e^n satisfies the moment constraints, but it is easy to show that as $n \to \infty$, the limiting performance is $J(0)$. ■

V. Conclusions and Future Work

A general framework for expressing stochastic optimizations under partial information was presented and specialized to shortest path optimization. As part of this specialization, a parametrization for information capacity is offered that allowed us to efficiently perform information optimization as well as provide an analytic relationship between capacity and performance. We also showed how to efficiently reduce the complexity of information optimization while incurring a provable bound in performance loss.

Future work includes expanding our methods of information optimization to approaches that use limited information about the graph as a means to balance the amount of information required of the distribution and the complexity of the bound. We also seek to study the effect of spreading information over subsequent decision (i.e., the agent can request information as it traverses the graph). Finally, we are interested in expanding our treatment of optimizing performance under information to the case of having multiple agents traverse the graph simultaneously and cause congestion along the edges. Specifically, we seek to understand the utility of information as we vary the number of agents.

References

[1] S. Walley, H. Tan, and A. Viterbi, "Shortest path cost distribution in random graphs with positive integer edge costs," in *Twelfth Annual Joint Conference of the IEEE Computer and Communications Societies*, mar 1993, pp. 1023–1032.

[2] R. Hassan and E. Zemel, "On shortest paths in graphs with random weights," *MATHEMATICS OF OPERATIONS RESEARCH*, vol. 10, no. 4, pp. 557–564, 1985.

[3] 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.

[4] 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.

[5] R. Cogill, "Randomized load balancing with non-uniform task lengths," in *Forty-Fifth Annual Allerton Conference*, sep 2007.

[6] K. Isii, "On sharpness of tchebycheff-type inequalities," *Annals of the Institute of Statistical Mathematics*, vol. 14, no. 1, pp. 185–197, 1962.

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

[8] ——, "Tight bounds on expected order statistics," *Probab. Eng. Inf. Sci.*, vol. 20, no. 4, pp. 667–686, 2006.

[9] N. Papadatos, "Maximum variance of order statistics," *Annals of the Institute of Statistical Mathematics*, vol. 47, no. 1, pp. 185–193, January 1995. [Online]. Available: http://ideas.repec.org/a/spr/aistmt/v47y1995i1p185-193.html

[10] L. Vandenberghe, S. Boyd, and K. Comanor, "Generalized chebyshev bounds via semidefinite programming," *SIAM Rev.*, vol. 49, no. 1, pp. 52–64, 2007.

[11] T. Aven, "Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables," *Journal of applied probability*, vol. 22, no. 3, pp. 723–728, 1985.