Performance analysis, Optimization and Optics

Eitan Bachmat [*] Sveinung Erland [†]

ABSTRACT

We introduce some methods and concepts of optics into performance analysis and optimization

1. A RESULT IN OPTICS

Fermat's principle in optics states that light will travel from a point A to a point B along a path that minimizes (locally) the amount of travel time. If the media surrounding A and B is homogeneous then travel time is proportional to length and light will take the minimal length path, i.e., a straight line (geodesic) between the points. More generally, the speed of light depends on the media in which light travels and each material has an index of refraction which is inversely proportional to the propagation speed of light in that material. As a result, when light travels in a certain media, it does so in straight lines, but at the boundary between different media it breaks (changes direction) in accordance with Fermat's principle (Snell's law). Mathematically, consider a path in 2 dimensions in which light always moves forward in the x direction. In that case, we can take x to be the path parameter, i.e. the path is given by $(x, y(x))$, say between $x = a$ and $x = b$. The time it takes light to traverse the path is given by $T(y) = \int_a^b ds$ where $(ds)^2 = I^2(x, y(x))((dx)^2 + (dy)^2)$, and $I(x, y)$ is the index of refraction of the media at location (x, y). Fermat's principle states that among all possible paths , light will choose a path that minimizes the expression $T(y)$ among all nearby paths.

Consider the case where the index of refraction I depends only on the x coordinate, i.e., has the form $I(x)$. An interval exchange transformation σ is a piecewise continuous (isometric) map from an interval $[a, b]$ to itself, which is obtained by subdividing $[a, b]$ into half open sub-intervals $J_1, J_2, \ldots, J_K$ and permuting their order (shifting the sub-intervals) by some permutation $\tilde{\sigma}$ to obtain a new subdivision. We also (rather arbitrarily) map b to itself. The map σ simply follows where each point in the interval goes. for example, if $[0, 1]$ is divided into $J_1 = [0, 1/6)$, $J_2 = [1/6, 1/2]$, $J_3 = [1/2, 1)$ and $\tilde{\sigma}$is the cyclic permutation $1 \to 2 \to 3 \to 1$, then for $x \in J_1$, $\sigma(x) = x + 1/2$, for $x \in J_2$, $\sigma(x) = x + 1/2$, for $x \in J_3$, $\sigma(x) = x - 1/2$ and $\sigma(1) = 1$. For a given interval exchange transformation σ, we can consider the associated index of refraction function $I_\sigma(x) = I(\sigma(x))$. We will also define $(ds_\sigma)^2 = I_\sigma^2(x)((dx)^2 + (dy)^2)$ and $T_\sigma(y) = \int_a^b ds_\sigma$.

THEOREM 1. *Fix a starting point $A = (a, y_1)$ and an ending point $B = (b, y_2)$ for paths (possible light trajectories). Assume ds has the form $(ds)^2 = I^2(x)((dx)^2 + (dy)^2)$, then, for any interval exchange transformation σ and any path $y(x)$ between A and B, there is a path $y_\sigma(x)$ which is constructed canonically from $y(x)$and σ such that $T(y) = T_\sigma(y_\sigma)$. In particular, the time it takes light to travel from A to B is the same for the index of refraction profiles $I(x)$ and $I_\sigma(x)$.*

Proof: We extend the transformation σ to the plane by $\sigma(x, y) = (\sigma(x), y)$. The map σ sends ds to ds_σ, since it send $I(x)$ to $I_\sigma(x)$ and both dx and dy are invariant under σ since both are translation invariant and σ is piecewise composed of translations along the x coordinate. Consequently, if $\sigma(y)$ denotes the image of the path y under σ we have $T(y) = T_\sigma(\sigma(y))$. The problem is that $\sigma(y)$ is neither continuous, nor starts in A nor ends in B in general. However, we can fix these issues by piecewise translations along the y axis and these preserve ds_σ since $I_\sigma(x)$ does not depend on y. To fix $\sigma(y)$, we Let $J_i = [a_i, b_i)$, $i = 1, \ldots, K$ be the interval sub-division associated with σ and let $\Delta_i(y) = y(b_i) - y(a_i)$, $h_i = y(a_i) - y_1$ and $\Delta(y) = \sum_i \Delta_i = y(b) - y(a)$. We consider $\sigma(y)$ on the sub-interval division given by $\sigma(J_{\sigma^{-1}(1)}), \sigma(J_{\sigma^{-1}(2)}), \ldots, \sigma(J_{\sigma^{-1}(K)})$ which is the new sub-division of $[a, b]$ induced by σ. We denote the endpoints of the new sub-division by $a = c_1, c_2, \ldots c_K = b$, i.e., $\sigma(J_{\sigma^{-1}(j)}) = [c_j, c_{j+1})$. The function $\sigma(y)$ restricted to the intervals $\sigma(J_{\sigma^{-1}(j)})$ is continuous, being a translation along the x axis of y restricted to $J_{\sigma^{-1}(j)}$. We define on the interval $\sigma(J_{\sigma^{-1}(j)})$,

$$y_\sigma = \sigma(y) - (h_{\sigma^{-1}(i)} - \sum_{l=1}^{i-1} \Delta_{\sigma^{-1}(j)}).$$

It is easy to check inductively that the shifted pieces match at the endpoints c_i and thus form a continuous map. In addition $\Delta(y) = \Delta(y_\sigma)$ as sums of permuted translation invariant coordinate differences. Consequently, $y_\sigma(b) = y_2$ since

$$y_\sigma(b) = y_\sigma(a) + \Delta(y_\sigma) = y(a) + \Delta(y) = y_1 + (y_2 - y_1) = y_2$$

[*] Eitan Bachmat, Department of Computer science, Ben-Gurion University, Beer-Sheva, Israel, 84105. ebachmat@cs.bgu.ac.il

[†] Sveinung Erland, Department of Maritime Studies, Western Norway University of Applied Sciences, 5528 Haugesund, Norway. Sveinung.Erland@hvl.no

and so y_σ is also a path from A to B. *q.e.d.*

2. A RESULT ON SITA QUEUES

SITA queues were introduced in [5] as a means for reducing job size variance at individual hosts in a multi-server setting. In essence, SITA queues generalize the basic idea of express line queues and are defined by a set of cutoffs $0 = s_0 < s_1 < s_2 < ... < s_{h-1} < s_h = \infty$, where h is the number of servers and the hosts are numbered (indexed) by $1, ..., h$. In general the hosts do not have to have the same speeds and we denote by c_i the speed of host i. This means that a job of size x takes x/c_i on host i, with c_1 normalized to be the reference speed, i.e., $c_1 = 1$. Given the cutoffs a job of size x (with respect to the first host) will be assigned to host i such that $s_{i-1} \leq x < s_i$. We will assume Poisson arrivals, and that the job sizes are drawn from an i.i.d. generic bounded distribution X, supported on some interval $[1, p]$ and with a strictly positive piecewise smooth density f. We also assume a fixed rate of arrival λ. Our target function will be the average waiting time. We are interested in exploring the regime where the number of hosts $h \to \infty$.

We will use a server reciprocal speed profile function $I(x)$ to describe the (reciprocal) speed of servers as their number increases. We say $c_{1,h}, \dots c_{h,h}$ is a *compatible family of server reciprocal speeds* if there is a function $I(x)$ such that

$$c_{i,h} = I(i/h) \qquad (1)$$

Similarly, we will use increasing functions to describe a rule to produce cutoffs for an ever increasing number of hosts h, Let $y : [0, 1] \longrightarrow [1, p]$ be an increasing function with $y(0) = 1$ and $y(1) = p$. We will think of such functions y as providing a formula for choosing the cutoffs. We say that $1 < s_{1,h} < \dots < s_{h-1,h} < p$ is a *compatible family of cutoffs*, if there is a function y as above such that

$$s_{i,h} = y(i/h) \qquad (2)$$

Let $E(W)(I, y, X, h, \lambda)$ be the mean waiting time for a SITA system with generic job size distribution X, arrival rate λ, h hosts with speeds given by the profile function I as in (1) and cutoffs given by the function y as in (2). We say that y_{opt} is an *h-asymptotically optimal cutoff function* with respect to I and X, if for any other family of cutoffs y, for any λ and for any $\varepsilon > 0$ we have

$$E(W)(I, y_{opt}, X, h, \lambda)/E(W)(I, y, X_p, h, \lambda) < 1 + \varepsilon$$

for h large enough.

We say that two reciprocal speed profiles $I_1(x)$ and $I_2(x)$ have asymptotically the same SITA performance in the large h regime if for all λ, and X

$$1 - \varepsilon < \frac{E(W)(I_1, y_{1,opt}, X, h, \lambda)}{E(W)(I_2, y_{2,opt}, X, h, \lambda)} < 1 + \varepsilon$$

THEOREM 2. *Let $I(x)$ be a reciprocal speed profile, σ an interval exchange transformation as above, then $I(x)$ and $I_\sigma(x)$ have asymptotically the same SITA performance in the large h regime.*

Proof: According to [1, 3] the average waiting time $E(W)(I, y, X, h, \lambda)$ is (asymptotically) proportional to $\int_0^1 ds$ with $ds = I(x)(yf(y))^2(y')^2dx = I(x)(yf(y))^2\frac{(dy)^2}{dx}$, where f is the density of the distribution X. Let $B(y) = (yf(y))^2$. We will show that we can find a change of coordinates $\tilde{x} = x$, $\tilde{y} = g(y)$ such that $\int_0^1 ds = \int_0^1 \tilde{ds}$ with $\tilde{ds} = I(x)\frac{(d\tilde{y})^2}{d\tilde{x}}$ The argument will be local at a given point (x, y) and we will consider a change of variables that locally looks like $\tilde{(y)} = ky$, i.e., $g'(y) = k$. for some constant k. It is convenient to think of y' as $\frac{dy}{dx}$ and since $d\tilde{y} = g'(y)dy$ we have

$$I(x)B(y)(y')^2dx = I(x)B(y)\frac{dy^2}{dx}$$

$$= I(x)B(y)\frac{d\tilde{y}^2}{g'(y)^2dx}$$

$$= I(x)B(y)\frac{(\tilde{y}')^2}{g'(y)^2}dx.$$

Choosing g to satisfy $g'(y) = \sqrt{B(y)}$ will eliminate the y dependence of the expression as required. Applying the same procedure with $I_\sigma(x)$ and the same job size distribution X, we note that the transformation g did not depend on I (or I_σ), just on B, and B depends just on the distribution X. Consequently, the same coordinate change yields in this case the expression $\tilde{ds}_\sigma = I_\sigma(x)\frac{(d\tilde{y})^2}{d\tilde{x}}$. The proof of our first theorem, now applies word for word. *q.e.d.*

3. LENSES AND OPTIMIZATION

Lenses form a central concept in Optics. The defining property of a (focal) lens is that there are (infinitely) many paths which solve Fermat's principle, i.e., minimize $\int ds$. We explain how a similar concept arises in certain resource allocation problems in the context of projects.

Classically, in Operations Research, a project is described via a vertex weighted directed acyclic graph. The vertices of the graph represent tasks, the weights represent the time required to complete the tasks. The directed edges represent precedence relations between the tasks. The end vertex of an edge is a task which cannot begin before the initial vertex task is completed. In this setting, every path through the directed graph can be given a weight which is the completion time of that particular sequence of tasks if they are executed one after the other with no delays. It is then shown that the The completion time of the project is given by the heaviest weight path in the graph and such a path is known as a critical path, since any further delay along this path will extend the project completion time. The analysis of the project can be done using the Critical Path Method (CPM) which is a dynamic programming procedure.

The problem of resource allocation comes in many flavors, but generally speaking, the idea is to shorten the project completion time by adding resources to tasks. As an example, let us assume that the completion times of the tasks (weights) can have just two values, say 1 and 2. We assume that initially all tasks are slow, i.e., take 2 time units. We further assume that by applying a fix amount of extra resources a task can be converted from a slow task to a fast task, requiring only a single time unit. Suppose further that we are provided with some (integer) target project completion time T. Our optimization challenge is to achieve the desired project completion time using as few resources as possible, i.e., by converting as few slow tasks as possible into fast tasks . A policy is the choice of which tasks to expedite by using the resources, namely, the resulting list of fast tasks. If the project completion time of a policy is at most T then we say that the policy is admissible. An optimal policy is an admissible one with a minimal number of fast tasks. An admissible policy will be called locally optimal if it does not strictly contain another admissible policy. We have the following trivial but basic observation.

claim: In an optimal, or more generally, in a locally optimal policy all the fast (weight 1) tasks are on a critical path.

To establish the claim, note that if a fast task v is not on any critical path, then the project completion time is at least 1 larger than the weight of the heaviest path through v. Converting v back to a slow task by not spending resources on it can only increase the weight of any given path by 1 so the project still completes on time with less resources contradicting optimality. *q.e.d.*

From the claim we see that optimal resource allocation policies are abstract lenses. When the tasks can be represented geometrically and the precedence relations can also be interpreted geometrically then optimal resource allocation policies will take (literally) the shape of a lens in that geometry.A particularly interesting case, which has surprising connections to SITA queues is airplane boarding, [2]. In airplane boarding, the tasks correspond to passengers, the task of each passenger is to sit down and clear the aisle. the project is complete when all tasks complete. Passengers/tasks can be represented geometrically in terms of coordinates (q, r) where q reflects queue position and r, the row number (normalized so that $0 \leq q, r \leq 1$).

In a non-congested airplane, a passenger (q_1, r_1) blocks another passenger at (q_2, r_2) if he/she is in front in the queue $q_1 \leq q_2$ and has a lower row number $r_1 \leq r_2$ providing a geometric interpretation of the precedence/blocking relations. Finally, different groups of passengers have different aisle clearing time distributions. For simplicity we will assume that a passenger belong to the first group takes 1 time unit to get organized, sit down and clear the aisle, while a passenger from the second group takes 2 time units to do the same. Our policy problem, is to find where should we place, both in the queue and in the airplane the group of slow passengers in order to minimize boarding time, i.e., which coordinates should slow passengers have in order to minimize their effect on boarding time. We are interested in asymptotic solutions when the number of passengers is large (goes to infinity).

Theorem 3. *If the group of slow passengers comprises at most 26% of all passengers, there is a policy so that asymptotically, the slow passengers have no effect on the boarding time, i.e., it offers the same boarding time as if all passengers were fast.*

idea of proof: We first note that the relevant geometry is that of Minkowski space, rather than Euclidean space as in classical optics and is given by $ds = I(q, r)\sqrt{dqdr}$, where $I(q, r) = 1$ for the usual media (fast passengers) and $I(q, r) = 2$ for the lens media. The analogue of Fermat's principle in Minkowski space are paths that maximize $\int ds$. We would like the lens to occupy as much space as possible (more slow passengers) but to also satisfy $max \int ds = max \int \tilde{ds}$ with $\tilde{ds} = \sqrt{dqdr}$ corresponding to the case where $I(q, r)$ is always 1 (all passengers are fast). We start with Descartes' construction of oval lenses for the line $q + r = 1$ adapted to the Minkowski geometry case (same polynomial equation but different locus of solutions). As in the Euclidean case, this provides a solution in a certain range, but in the Minkowski case, we can optimize further (add lens material) by welding Descartes' curve to a semi-algebraic curve that creates a lateral wave that focuses on the endpoints of Descartes' curve. It then remains to compute the area and check that its more than 26%. In formulas (for a general speed ratio T),

$$\alpha_T = (T - \sqrt{T^2 - 1})^2 \tag{3}$$

Let q_T be given by

$$q_T = \frac{T - \sqrt{(T-1)(T+1)}}{2(T+1)} \tag{4}$$

Let c_T satisfy the equation

$$c_T = \alpha_T^{-1} x_T^{1-\alpha_T} \tag{5}$$

Let

$$z_T(q) = c_T q^{\alpha_T} \tag{6}$$

Let

$$r_T(q) = (\frac{1}{T^2}[T(T-1)+(2-T^2)q+2\sqrt{(1-T^2)q^2 + T(T-1)q}] \tag{7}$$

We define the curve S as $(x, z_T(q))$ for $0 \leq q \leq q_T$, and $(x, r_T(q))$ for $q_T \leq q \leq 1/2$.

Let I_d denote the reflection through the line $q = r$ and I_a the reflection through the line $r = 1 - q$. Let $\tilde{D}_T$ be the domain, bounded from above by the curve $r = 1 - q$, $0 \leq q \leq 1$ and from below by the curve S. let D_T be the domain which is the union of $\tilde{D}_T, I_d(\tilde{D}_T), I_a(\tilde{D}_T), I_a(I_d(\tilde{D}_T))$. Let $I(q, r) = T$ iff $(q, r) \in D_T$ and $I(q, r) = 1$ otherwise. Set $T = 2$ to get the required solution. *q.e.d.*

For non-congested airplanes, we conjecture that this construction is optimal in the sense that there is no critical lens with larger area. For congested airplanes, as congestion becomes large, the optimum approaches 25% and the shape of the optimal lens is very different (places slow passengers mainly in the back of the airplane). See [4] for more optics based analysis of airplane boarding.

Acknowledgments: The work of Eitan Bachmat was supported by the German Science Foundation (DFG) through grant JA 2311/3-1.

4. REFERENCES

[1] J. Anselmi and J. Doncel, Asymptotically Optimal Size-Interval Task Assignments, IEEE Transactions on parallel and distributed systems, vol. 30, no. 11, November 2019.

[2] E. Bachmat, Airplane boarding meets express line queues, European Journal of Operational Research Volume 275, Issue 3, 16 June 2019, Pages 1165-1177.

[3] E. Bachmat and A. Natanzon, Analysis of SITA queues with many servers and spacetime geometry, Perform. Evaluation Rev. 40(3): 92-94 (2012).

[4] S. Erland, J. Kaupuzs, V. Frette, R. Pugatch, and E. Bachmat, Lorentzian-geometry-based analysis of airplane boarding policies highlights slow passengers first as better, Phys. Rev. E 100, 062313, 2019.

[5] M. Harchol-Balter, M. Crovella and C. Murta, On choosing a task assignment policy for a distributed server system, *IEEE Journal of parallel and distributed computing*, Vol. 59, 204-228, 1999.