

UC Berkeley

UC Berkeley Previously Published Works

Title

Efficiency of ETA Prediction

Permalink

<https://escholarship.org/uc/item/08h8c0vn>

Journal

SIAM Journal on Mathematics of Data Science, 6(2)

ISSN

2577-0187

Authors

Yan, Chiwei

Johndrow, James

Woodard, Dawn

et al.

Publication Date

2024-06-30

DOI

10.1137/23m155699x

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Efficiency of ETA Prediction

Chiwei Yan* James Johndrow† Dawn Woodard‡ Yanwei Sun§

May 8, 2024

Abstract

Modern mobile applications such as navigation services and ride-sharing platforms rely heavily on geospatial technologies, most critically predictions of the time required for a vehicle to traverse a particular route, or the so-called estimated time of arrival (ETA). There are various methods used in practice, which differ in terms of the geographic granularity at which the predictive model is trained — e.g., segment-based methods predict travel time at the level of road segments (or a combination of several adjacent road segments) and then aggregate across the route, whereas route-based methods use generic information about the trip, such as origin and destination, to predict travel time. Though various forms of these methods have been developed, there has been no rigorous theoretical comparison regarding their accuracies, and empirical studies have, in many cases, drawn opposite conclusions. We provide the first theoretical analysis of the predictive accuracy of various ETA prediction methods and argue that maintaining a segment-level architecture in predicting travel time is often of first-order importance. Our work highlights that the accuracy of ETA prediction is driven not just by the sophistication of the model but also by the spatial granularity at which those methods are applied.

1 Introduction

Geospatial (maps) technologies underlie a broad spectrum of modern mobile applications. For example, consumer-facing navigation applications (such as Google Maps and Waze) provide recommended routes along with associated times, as well as turn-by-turn navigation along those routes. Geospatial technologies are also the foundation of decision systems for ride-sharing (such as Uber, Lyft, Didi Chuxing, and Ola) and delivery platforms (such as Uber Eats and Doordash). For example, riders on these platforms are presented with estimated pickup time and time to arrival, and drivers are provided with turn-by-turn navigation. Matching and pricing decisions on these platforms also heavily rely on mapping inputs to optimize efficiency and reliability [Yan et al., 2020].

An important geospatial technology is the prediction of the time required for a driver (or biker or pedestrian) to travel a particular route in the road network or the so-called estimated time of arrival (ETA). Modern methods leverage location data traces from past vehicle trips in the road network, so-called “floating-car” data, typically gathered (with permission) from users of a particular application, such as a consumer-facing navigation service. Location traces from driver trips in the road network are processed by a “map-matching” algorithm to obtain travel time

*Department of Industrial Engineering and Operations Research, University of California, Berkeley.

†Department of Statistics and Data Science, Wharton School, University of Pennsylvania.

‡LinkedIn, work done while at Uber Technologies.

§Department of Analytics, Marketing and Operations, Imperial College Business School.

observations on each road segment along the driver’s trajectory [Quddus et al., 2007]. This data provides detailed information about traffic and travel speed patterns throughout the road network and along individual routes, and so is the foundation of modern methods for ETA prediction at scale. A feature that distinguishes different prediction methods is *the level of geographic granularity* at which the predictive model is trained. The geographic unit can be a single road segment, a combination of multiple adjacent and connecting segments (aka “super-segment”, see Derrow-Pinion et al. 2021 for the implementation in Google Maps), or the entire trip. To be more specific, *segment-based* methods rely heavily on the underlying road network and predict travel time at the level of road segments or super-segments, and then aggregate across the route (see, e.g., Hofleitner et al. 2012 and Jenelius and Koutsopoulos 2013). On the other hand, with a large and growing amount of trip data being collected by firms such as ride-sharing platforms, a class of more recently proposed *route-based* methods hinge less on the road network and use generic information about the origin, destination, departure time and sometimes route characteristics to predict travel time. This started with the k -nearest neighbors approach proposed in Wang et al. [2016], where the prediction of travel time on a new route was done using travel times of historical trips that have similar origins, destinations and departure times as those of the predicting route. Then a number of neural-network based approaches were developed for route-based prediction [Jindal et al., 2017, Li et al., 2018, Yuan et al., 2020], including work from Didi Chuxing, a major ride-sharing provider.

Though many variations of these methods have been proposed in the literature and used in practice, there is a very limited theoretical understanding of the accuracy of these methods. Most work in this space is empirical, and these empirical studies have, in many cases, drawn opposite conclusions (see, e.g., Wang et al. 2018a, Yuan et al. 2020, Derrow-Pinion et al. 2021, Wang et al. 2016). Indeed, the comparison is not trivial. Segment-based methods have the advantage of a larger sample size as there are more individual traversals on a segment level. However, the estimation can accumulate errors due to aggregating over road segments. On the other hand, route-based methods can have the advantage of absorbing errors among segment travel times, but it is often at the cost of a smaller sample size. Part of the confusion in the empirical analyses stems from some papers assuming that the route that the driver will take is known, whereas other papers assume that the route is uncertain (and so must be estimated or ignored); the latter naturally disadvantages segment-based methods. However, even papers that analyze cases where the route is known sometimes conclude that route-based methods are superior [Wang et al., 2016]. Due to the uncertainty about the best approach to take, several recent papers have tried combining segment-based and route-based methods into a single model [Wang et al., 2018b, Hu et al., 2022], or using methods that model travel time at the level of the “super-segment” (a sequence of segments) [Wang et al., 2014, Derrow-Pinion et al., 2021].

To fill this gap in understanding, we conduct rigorous analyses comparing segment-based and route-based methods in terms of their predictive accuracy as a function of the training data sample size, i.e., in terms of statistical efficiency. We now give a brief summary of our framework, analysis, and major results.

Framework. We consider a road network consisting of a set of road segments (directed edges) between intersections (vertices). The training data consists of individual trips traversed at different times, each of which is along a route that consists of a sequence of adjacent road segments. The travel time on each road segment of that route is observed, and the total travel time of a trip is the sum of the observed segment travel times along the route. Our goal is to predict the total travel time on a new trip, given the dataset of historical trips. We ask questions such as: What is the most accurate ETA prediction method? and, How do various methods used in practice compare? To precisely answer these questions, we construct a data-generating process based on a

general mean function that depends flexibly on features like distance and time of day, and that also incorporates parameters associated with the idiosyncratic travel speed effects of individual road segments. Travel times are allowed to be correlated across the road segments of a trip. Under this data-generating process, we first explicitly characterize the optimal predictor that has the lowest predictive mean squared error. The optimal predictor, though having the best accuracy, is computationally intractable to implement in real-time mapping services and requires full knowledge of the segment travel times’ covariance structure which can be hard to obtain in practice. This calls for the need to understand the accuracy of simpler and more practical methods. We formally define a family of segment-based methods and route-based methods that resemble many practical methods proposed in the literature and used in practice.

Analysis and Results. We start with a finite-sample setting where a set of historical trips are given on an arbitrary road network. When segment travel times are *non-negatively* correlated over the network, we show that the predictive mean squared error of the optimal segment-based method, where prediction is made on each individual road segment and then aggregated over the predicting route, is always lower than those of a wide range of route-based methods. We then extend our analysis to an asymptotic setting where the number of trip observations grows with the size of the road network, and trip routes are sampled randomly from a generic route distribution. We show that a very simple class of segment-based methods with minimum information requirement can asymptotically dominate popular route-based methods. Furthermore, under a broad range of trip-generating processes on a grid network, we show that this class of simple segment-based methods is at least as good (up to a logarithmic factor) as any possible predictor. In other words, segment-based methods are asymptotically optimal up to a logarithmic factor. Numerical experiments based on realistic parameters reveal that the accuracy of the segment-based methods is extremely competitive — the error of the segment-based method is often very close to that of the optimal method, even not at the asymptotic limit.

Our analysis is greatly facilitated by the fact that minimizing the predictive mean squared error is mathematically equivalent to minimizing the estimation error for the expected travel time (i.e., the conditional mean of travel time given the input features). This allows us to focus our analysis on the accuracy of estimating the expected travel time (rather than, for example, the variance of travel time), which is summable across road segments. We believe that focusing on the accuracy of estimation of the expected travel time is reasonable because modern navigation applications have chosen to focus mainly on communicating the expected travel time to users, ignoring the variance (which is harder to communicate and less interpretable for many users).¹

In short, our paper makes a contribution to the literature and practice of the ETA prediction problem by providing important theoretical underpinnings. Through extensive analyses, our paper argues that maintaining a segment-level architecture in predicting travel time is often of first-order importance. This gives important practical guidance to mapping services as they improve their underlying predictive models. The remainder of the paper is organized as follows. In Section 2, we introduce the model setup and conduct finite sample analysis, meaning analysis for a given set of historical trips. In Section 3, we analyze an asymptotic setting in which the number of trip observations grows with the road network size, and trip observations are sampled randomly from a route distribution. We conclude with a brief discussion in Section 4. All proofs and various

¹As far as we know, currently there is only one case where a navigation provider gives a measure of variability of travel time to the user. This is for the web interface for Google Maps, in the case where the user inputs a future time for departure/arrival. In this case, Google Maps provides a range (interval prediction) for travel time. To our knowledge this information isn’t provided in the Google Maps app, or by any other common mapping providers like Apple Maps.

auxiliary results are presented in the supplement. A companion Jupyter notebook can be found at <https://github.com/yanchiwei/eta/blob/main/examples.ipynb> to reproduce all the examples presented in the paper.

2 Model and Finite Sample Analysis

We first consider a standard travel time setting, where we are given N historical trips on an arbitrary road network $(\mathcal{V}, \mathcal{S})$ where \mathcal{V} is a vertex set and \mathcal{S} is an edge (road segment) set. Let y_1, \dots, y_N be the routes for each trip, and $[N] := \{1, \dots, N\}$, so that $y_{[N]}$ is the set of routes. Put $|y_n|$ as the number of segments on route n . Each route consists of a sequence of distinct road segments $s \in \mathcal{S}$. Most simply, think of a road segment as the primitive used in the standard representation of road graphs, i.e., a directed section of roadway that is uninterrupted by intersections and has constant values for features like the number of lanes and speed limit. A more sophisticated representation of road graphs also fits into our framework: one which incorporates turn effects by defining a road segment s to be a section of roadway (with constant feature values) that is followed by a specific turn direction. For example, segment s can represent a particular directed section of highway that is followed by the turn onto an exit ramp, and the next segment s' in the route could be the exit ramp that is followed by a left turn onto a minor road (see e.g., Section 4.1 of Delling et al. [2017]). Let $T_{n,s}$ be the travel time on segment $s \in y_n$ for the n^{th} observed trip, and denote the n^{th} trip by $\mathcal{T}_n = \{y_n, \{T_{n,s}\}_{s \in y_n}\}$.

2.1 Generative Process

We first discuss the generative process that we assume for travel times. In practice, the segment travel time $T_{n,s}$ and the route travel time $\sum_{s \in y_n} T_{n,s}$ are affected by the set of observed features V_s of the road segments such as the number of lanes, speed limit, segment length, and road classification (local road, highway, arterial, etc.). The travel times are also affected by a set of trip-level characteristics W_n , such as time of week and weather conditions. In addition, there are unobserved idiosyncratic characteristics of the road segments that affect their travel times. For example, some segments have bad traffic conditions, a poor layout of the lanes, road constructions, or a slow traffic light, which the mapping services typically don't observe directly outside of the location trace data. Following this physical understanding, we assume that the segment travel times $T_{n,s} = g(\theta_s, V_s, W_n) + \varepsilon_{n,s}$, where $g(\theta_s, V_s, W_n)$ is the true mean with some function $g(\cdot)$, θ_s is an unobserved feature vector for each road segment s , and $\varepsilon_{n,s}$ is the error term with mean 0. Let $\theta := [\theta_s]_{s \in \mathcal{S}}$. The mean of the travel time on route y_n is then $\sum_{s \in y_n} g(\theta_s, V_s, W_n)$. For mathematical tractability, we analyze a simplified generative model that has an additive structure, i.e., we assume that $g(\theta_s, V_s, W_n) = \theta_s + h(V_s, W_n)$ for some function $h(\cdot)$ and for θ_s a scalar that can capture the fact that a particular road segment s has faster or slower average travel time. This generative model, while simple, captures the most foundational characteristics of typical traffic data, specifically a mean structure that depends in a potentially nonlinear way on W_n and V_s , as well as idiosyncratic travel time effects at the level of the road segment. Such additive models are common in the statistics literature, where they are called mixed-effects models Pinheiro and Bates [2006]. These discussions lead us to the following assumptions regarding the generative process of $T_{n,s}$.

Assumption 1. We make the following assumptions about $T_{n,s}$,

1. $T_{n,s} = \theta_s + h(V_s, W_n) + \varepsilon_{n,s}$ for some function h of the input features V_s, W_n , and for θ_s a scalar capturing road segment travel time effects.

2. For every trip n , the errors $\{\varepsilon_{n,s}\}_{s \in y_n}$ are drawn from a joint distribution with mean 0 for all $\varepsilon_{n,s}$ and covariances $\{\sigma_{s,t}\}_{s,t \in y_n}$, where $\sigma_{s,s} = \sigma_s^2$ is the variance of the error term on segment s .
3. For any $n \neq n'$ and any $s \in y_n, t \in y_{n'}$, $\varepsilon_{n,s}$ and $\varepsilon_{n',t}$ are independent.

The first and second assumptions are directly motivated by the discussions above. Note that we do not impose any distributional assumptions other than specifying the means and covariances of the travel times $T_{n,s}$. The third assumption says that conditional on all the segment-level and trip-level effects, the travel times on different trips are independent. This is a natural assumption, since much of the observed correlation across trips is due to time of week and other covariates. Conditional on those relevant covariates, it is much more reasonable to assume independence. Empirical evidence also shows that intra-trip correlation is much stronger than inter-trip correlation within similar time of week (see Figure 5 in Woodard et al. 2017).

2.2 Travel Time Estimators

For a new $(N+1)^{\text{th}}$ trip $\mathcal{T}_{N+1} = \{y_{N+1}, \{T_{N+1,s}\}_{s \in y_{N+1}}\}$ with segment-level feature sets $\{V_s\}_{s \in y_{N+1}}$ and route-level feature set W_{N+1} , the goal is to come up with an estimator $\hat{\Theta}_{\mathcal{T}_{N+1}}$, a function of the N historical trips, $\{\mathcal{T}_n\}_{n \in [N]}$, for the total travel time $\sum_{s \in y_{N+1}} T_{N+1,s}$ that minimizes the following *predictive mean squared error* where the expectations are taken over $\{T_{n,s}\}_{n \in [N+1], s \in y_n}$ conditional on h and on $\{\theta_s, V_s, W_n\}_{s \in y_n, n \in [N+1]}$. We drop the explicit conditioning in the following expectations for notation brevity.

$$\begin{aligned} & \mathbb{E} \left[\left(\hat{\Theta}_{\mathcal{T}_{N+1}} - \sum_{s \in y_{N+1}} T_{N+1,s} \right)^2 \right] \\ = & \mathbb{E} \left[\left(\hat{\Theta}_{\mathcal{T}_{N+1}} - \sum_{s \in y_{N+1}} (\theta_s + h(V_s, W_{N+1})) + \sum_{s \in y_{N+1}} (\theta_s + h(V_s, W_{N+1})) - \sum_{s \in y_{N+1}} T_{N+1,s} \right)^2 \right] \\ = & \mathbb{E} \left[\left(\hat{\Theta}_{\mathcal{T}_{N+1}} - \sum_{s \in y_{N+1}} (\theta_s + h(V_s, W_{N+1})) \right)^2 \right] + \mathbb{E} \left[\left(\sum_{s \in y_{N+1}} (\theta_s + h(V_s, W_{N+1})) - \sum_{s \in y_{N+1}} T_{N+1,s} \right)^2 \right]. \end{aligned}$$

The last equality holds because $\hat{\Theta}_{\mathcal{T}_{N+1}}$ (a function of $\{T_{n,s}\}_{s \in y_n, n \in [N]}$) and $\sum_{s \in y_{N+1}} T_{N+1,s}$ are independent conditional on $\{\theta_s, V_s, W_n\}_{s \in y_n, n \in [N+1]}$ by the third part in Assumption 1, and moreover $\mathbb{E}[\sum_{s \in y_{N+1}} T_{N+1,s}] = \sum_{s \in y_{N+1}} (\theta_s + h(V_s, W_{N+1}))$. Now notice that the second term in the last equality does not depend on $\hat{\Theta}_{\mathcal{T}_{N+1}}$. This implies that the estimator $\hat{\Theta}_{\mathcal{T}_{N+1}}$ that minimizes the predictive error is the same one that minimizes the squared error for estimating the mean $\sum_{s \in y_{N+1}} (\theta_s + h(V_s, W_{N+1}))$.

Segment-based approaches typically directly estimate segment-level embeddings θ_s as well as a mean function $g(\cdot)$, and then they sum up the estimated travel times across the segments of the route. For example, the production ETA model in Google Maps at the time of the publication of DeepMind [2020], Derrow-Pinion et al. [2021] was based on a linear regression model for $g(\cdot)$ with features V_s that include length, road class, and real-time and historical average travel speed; it also included learnable embedding vectors θ_s for each road segment to capture idiosyncratic effects. For tractability, we analyze a class of segment-based models that fit into the additive framework where $T_{n,s} = \theta_s + h(V_s, W_n) + \varepsilon_{n,s}$ for a scalar θ_s .

Route-based methods, unlike segment-based methods, fit a model for the whole trip travel time $\sum_{s \in y_n} T_{n,s}$. Typically they include some embeddings at the level of the origin and destination,

or at the level of origin-destination pair. They also typically use trip-level features W_n , as well as route-level features \tilde{V}_n that are created by aggregating over the segments of the route, such as route length. An example is a deep neural network used by Uber [Hu et al., 2022], which includes an embedding $\Theta_{o,d}$ for the origin-destination pair, as well as trip-level features W_n that include time of week, and route features \tilde{V}_n that include route length and aggregated inputs related to real-time traffic conditions. In our simplified setting, a route-based method corresponds to fitting a model $\Theta_{y_n} + f(\tilde{V}_n, W_n)$, where Θ_{y_n} is an embedding that approximates $\sum_{s \in y_n} \theta_s$ and $f(\tilde{V}_n, W_n)$ is a function that approximates $\sum_{s \in y_n} h(V_s, W_n)$ by using a feature transformation vector $\tilde{V}_n = \phi(\{V_s\}_{s \in y_n})$. For example, if V_s is the length of segment s , \tilde{V}_n can be the total length of the route y_n (see, e.g., the linear regression model based on trip distance on page 8 of Wang et al. 2016).

In most of the travel time models described in the literature, the parameters in the function $h(\cdot)$ or $f(\cdot)$ don't need to scale with the number of road segments of the network (because the set of features, which include things like segment/route length and weather conditions, is fixed). The number of parameters $\{\theta_s\}_{s \in \mathcal{S}}$, by contrast, scales proportionally with the number of road segments. Since typical road networks, for example for large metropolitan regions, have hundreds of thousands or even millions of road segments, the number of parameters in the set $\{\theta_s\}_{s \in \mathcal{S}}$ tends to dominate the parameter size needed to robustly model $h(\cdot)$ or $f(\cdot)$. For example, the linear regression production model described in DeepMind [2020], Derrow-Pinion et al. [2021] estimates the parameters of a regression model with fixed input and output dimension. This model also includes road segment embeddings, the number of which scales proportionally with the size of the road graph.

As a result, most of the error in the estimation of $(\theta_s + h(V_s, W_n))$ in a segment-based method typically comes from the error in the estimation of θ_s . A similar effect occurs in route-based methods: so long as $\phi(\cdot)$ is chosen in such a way that $f(\tilde{V}_n, W_n)$ is a good approximation to $\sum_{s \in y_n} h(V_s, W_n)$, typically it is much easier to estimate $f(\cdot)$ than to estimate Θ_y . If $\phi(\cdot)$ is chosen in such a way that $f(\tilde{V}_n, W_n)$ is *not* a good approximation to $\sum_{s \in y_n} h(V_s, W_n)$, then the accuracy of the route-based method is degraded. We assume that this is not the case, which gives the benefit of the doubt to the route-based method.

Based on the discussions above, we will thus focus on estimating the accumulation of segment random effects on a route, $\sum_{s \in y_n} \theta_s$. We denote by $T'_{n,s} = T_{n,s} - h(V_s, W_n)$ the adjusted observed segment travel time with mean θ_s .

Assumption 2. We assume that the function $h(\cdot)$ is known (approximating a situation where $h(\cdot)$ is much easier to estimate than $\{\theta_s\}_{s \in \mathcal{S}}$).

To compare the predictive accuracy of different estimators, we introduce the *integrated risk*, a Bayesian statistical concept capturing the accuracy of the travel time prediction by integrating the risk (in our case, the predictive mean squared error) over the prior distribution of the unknown parameters. This is also known as an ‘‘average-case’’ analysis of accuracy (versus for example a worst-case analysis). As we shall see later, focusing on such an average-case analysis also helps us to reach more general conclusions regarding the comparisons of these estimators. In particular, we impose the following assumption on the prior distribution.

Assumption 3. We assume that $\{\theta_s\}_{s \in \mathcal{S}}$ are drawn i.i.d. from a population distribution with mean μ and variance τ^2 .

The choice of i.i.d. population distribution is for notation brevity, and our results can be generalized to non-i.i.d. population distribution to capture, for example, congestion patterns across

road networks. The integrated risk of estimator $\hat{\Theta}_y$ for a new route y given historical routes $y_{[N]}$, which we call $R(\hat{\Theta}_y | y_{[N]})$, is defined to be the expectation of the squared difference between the true total mean travel time $\Theta_y := \sum_{s \in y} \theta_s$ and the estimated total mean travel time $\hat{\Theta}_y$. This expectation is taken with respect to (i) the observed adjusted segment travel times $\{T'_{n,s}\}_{n \in [N], s \in y_n}$ and (ii) the population distribution over the parameters $\{\theta_s\}_{s \in y}$, conditional on the historical route observations $y_{[N]}$:

$$R(\hat{\Theta}_y | y_{[N]}) := \mathbb{E} \left[\left(\hat{\Theta}_y - \Theta_y \right)^2 \middle| y_{[N]} \right]. \quad (1)$$

We now illustrate how travel time on a route can be predicted using different examples of estimators.

Example 1 (TRAVEL TIME ESTIMATORS). Consider the following 3×3 grid in Figure 1 where there are six historical trips $\{\mathcal{T}_n\}_{n \in \{1, \dots, 6\}}$ whose routes are displayed in the figure. Let segments s_1 and s_2 denote the ones that traverse $(1, 0) \rightarrow (1, 1)$ and $(1, 1) \rightarrow (1, 2)$, respectively. Suppose that we want to predict the travel time on route y which traverses through $(1, 0) \rightarrow (1, 1) \rightarrow (1, 2)$ (s_1 and s_2). Among the space of all possible estimators which are functions of historical data, the following are a few simple estimators that capture characteristics of popular estimators used in the literature and practice.

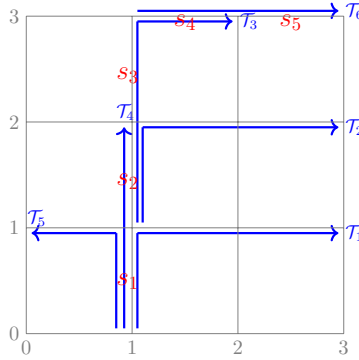


Figure 1: A 3×3 grid example.

1. *Segment-based estimator:* We estimate travel times on each segment using individual segment traversal data and then aggregate them across the route. In this example, s_1 has three traversals (from trips $\mathcal{T}_1, \mathcal{T}_4$ and \mathcal{T}_5), s_2 has three traversals (from trips $\mathcal{T}_2, \mathcal{T}_3$ and \mathcal{T}_4). If we use a simple estimator by taking the average of historical travel times,

$$\hat{\Theta}_y = \frac{T'_{1,s_1} + T'_{4,s_1} + T'_{5,s_1}}{3} + \frac{T'_{2,s_2} + T'_{3,s_2} + T'_{4,s_2}}{3},$$

where each term is the estimation of the segment travel times on s_1 and s_2 respectively.

2. *Generalized segment-based estimator:* Combining individual segments into *super-segments*, we can generalize the previously defined segment-based estimator. For example, we can define s_1 and s_2 together as one super-segment, and this super-segment has one traversal (from trips \mathcal{T}_4). This yields

$$\hat{\Theta}_y = T'_{4,s_1} + T'_{4,s_2}.$$

3. *Route-based estimator*: Instead of aggregating over segments or super-segments, we directly use total travel times on historical routes that are similar to route y for estimation. For example, we can average over travel times of routes that share similar origin and destination as y — both the origin and the destination of the route are at most one segment away from those of y . This includes trips \mathcal{T}_4 and \mathcal{T}_5 . This gives,

$$\hat{\Theta}_y = \frac{\sum_{s \in y_4} T'_{4,s} + \sum_{s \in y_5} T'_{5,s}}{2}.$$

Note that for route-based estimators, it is possible to use historical traversal data on segments that are *not* included in y to predict the total travel time of y .

Variations of the abovementioned estimators have been practiced extensively, but there is a lack of formal analysis to compare their accuracies. Such comparison is not straightforward because these estimators differ in sample sizes and the way historical data is used, and these differences can have non-trivial effects on accuracies, as we will illustrate later.

2.2.1 The Optimal Estimator under Normality Assumptions

Before comparing different estimators, it is always helpful to understand the perfect benchmark first — the optimal estimator $\hat{\Theta}_y^*$ that has the minimal integrated risk $R(\hat{\Theta}_y^* | y_{[N]})$. Characterizing the optimal estimator under general distributions does not always admit closed or tractable forms [Gelman et al., 2013]. Here we give the form of the optimal estimator under an additional assumption that the error terms $\varepsilon_{n,s}$ and θ_s are jointly Gaussian distributed. Defining $M = \sum_{n=1}^N |y_n|$, let $Z \in \mathbb{R}_{\geq 0}^{M \times 1}$ be the vector consisting of the concatenated travel times across all the trips so that Z_i is the travel time for a single segment on a single trip. Similarly, let $\mathcal{E} \in \mathbb{R}^{M \times 1}$ be the corresponding vector of error terms. Let $u_i = s$ if the i^{th} entry in Z and \mathcal{E} is a travel time and its corresponding error term on segment s . Let $w_i = n$ if the i^{th} entry in Z and \mathcal{E} is a travel time and its corresponding error term from trip n . Thus, $Z_i = T'_{w_i, u_i}$, $\forall i \in \{1, \dots, M\}$. Let $U \in \{0, 1\}^{M \times |\mathcal{S}|}$ be a matrix with entries $U_{i,s} = \mathbf{1}\{u_i = s\}$. This yields that $Z = U\theta + \mathcal{E}$.

Define $\Phi \in \mathbb{R}^{M \times M}$ to be a matrix whose entries are

$$\Phi_{i,j} = \begin{cases} \sigma_{u_i, u_j}, & \text{if } w_i = w_j, \\ 0, & \text{otherwise.} \end{cases}$$

Notice that Φ is a block diagonal matrix with N total blocks, where the n^{th} diagonal block has dimension $|y_n| \times |y_n|$. Let $e \in \mathbb{1}^{|\mathcal{S}| \times 1}$ be a $|\mathcal{S}|$ -dimensional all-ones vector and $e_y \in \{0, 1\}^{|\mathcal{S}| \times 1}$ such that $e_{y,s} = 1$ if $s \in y$ and 0 otherwise. To simplify the notation, put $E_y = e_y e_y^\top$ and $Q = U^\top \Phi^{-1} U + \text{diag}((1/\tau^2)e)$.

Theorem 1 (OPTIMAL ESTIMATOR). *In addition to Assumptions 1 and 3, assume that (\mathcal{E}, θ) are jointly Gaussian distributed, the following estimator $\hat{\Theta}_y^*$ of travel time on route y minimizes the integrated risk (1) among all possible estimators,*

$$\hat{\Theta}_y^* = e_y^\top Q^{-1} (U^\top \Phi^{-1} Z + (\mu/\tau^2)e). \quad (2)$$

Its integrated risk $R(\hat{\Theta}_y^ | y_{[N]})$ based on squared error is*

$$\text{tr}(\Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top) + \text{tr}(\text{diag}(\tau^2 e) (E_y + U^\top \Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top \Phi^{-1} U - 2E_y Q^{-1} U^\top \Phi^{-1} U)).$$

The first term is the expected variance, and the second term is the expected squared bias.

An interesting observation is that the minimum integrated risk does *not* depend on the population mean μ . The proof relies on deriving the conditional mean $\mathbb{E}[\sum_{s \in y} \theta_s | \{\mathcal{T}_n\}_{n \in [N]}]$ of travel time on route y given historical trip data. It then uses the fact that the estimator minimizing the integrated risk based on squared error is the posterior mean (see, e.g., Berger 2013a). As a sanity check, when there is no historical traversal at all, $\hat{\Theta}_y^* = |y|\mu$ which simply uses the population mean.

When segment travel times are independent across the road network, the form of the optimal estimator can be greatly simplified. Let $N_s := |\{y_n : s \in y_n, n \in [N]\}|$ denote the sample size of traversals on segment s in the historical data.

Corollary 1 (OPTIMAL ESTIMATOR UNDER INDEPENDENT SEGMENT TRAVEL TIMES). In addition to Assumptions 1 and 3, assume that (\mathcal{E}, θ) are jointly Gaussian distributed, when $\sigma_{s,t} = 0, \forall s \neq t \in \mathcal{S}$, the optimal estimator takes the following form,

$$\hat{\Theta}_y^* = \sum_{s \in y} \left(\frac{\sigma_s^2}{N_s \tau^2 + \sigma_s^2} \cdot \mu + \frac{N_s \tau^2}{N_s \tau^2 + \sigma_s^2} \cdot \frac{\sum_{n: s \in y_n} T'_{n,s}}{N_s} \right).$$

This result says that in the independent case, the optimal estimator $\hat{\Theta}_y^*$ takes the form of a simple segment-based estimator, where the travel times on each segment are estimated by a weighted average of the sample mean $\sum_{n: s \in y_n} T'_{n,s}/N_s$ and the population mean μ . The weight depends on the sample size N_s and variance parameters σ_s^2 and τ^2 . Intuitively, when the sample size N_s or population variance τ^2 is high, the sample mean gains more weight — the optimal estimator weighs the historical observations more; on the other hand, when the variance of the segment travel time σ_s^2 is high, the optimal estimator relies more on the population information.

When segment travel times are correlated over the road network, the form of the optimal estimator cannot be decomposed by segments, and computing the optimal estimator for a new route y uses *all* historical segment traversal data over the *entire* road network, regardless of being part of route y or not. Below is an example.

Example 2. Using the same setup in Example 1, we now compute the optimal estimator for travel time on route y , traversing through $(1, 0) \rightarrow (1, 1) \rightarrow (1, 2)$. Take the example where $\{\theta_s\}_{s \in \mathcal{S}}$ are drawn i.i.d. from a normal distribution with mean $\mu = 1$ and variance $\tau^2 = 0.2$ and the adjusted segment travel times are drawn from a multivariate normal distribution with mean $\{\theta_s\}_{s \in \mathcal{S}}$ and covariance matrix $\Sigma = [\sigma_{s,t}]_{s,t \in \mathcal{S}} = e^{-\mathcal{L}}$. Matrix \mathcal{L} is the normalized graph Laplacian of an undirected graph where each node in the graph is a segment in the 3×3 grid network in Figure 1, and an edge is created if two segments are directly connected. Precisely, let $\mathcal{L} = D^{-1/2} L D^{1/2}$ where D is the diagonal matrix of segment degrees, and $L = D - A$ is the graph Laplacian where A is the adjacency matrix. The covariance matrix is the matrix exponential $e^{-\mathcal{L}}$ which is often called the diffusion kernel. It models spatial decay of correlation among segment travel times. For this example, the optimal estimator $\hat{\Theta}_y^*$ based on (2) takes the form

$$\begin{aligned} \hat{\Theta}_y^* = & 0.211 \cdot T'_{1,(1,0) \rightarrow (1,1)} - 0.040 \cdot T'_{1,(1,1) \rightarrow (2,1)} + 0.002 \cdot T'_{1,(2,1) \rightarrow (3,1)} \\ & + 0.207 \cdot T'_{2,(1,1) \rightarrow (1,2)} - 0.003 \cdot T'_{2,(1,2) \rightarrow (2,2)} + 0.002 \cdot T'_{2,(2,2) \rightarrow (3,2)} \\ & + 0.210 \cdot T'_{3,(1,1) \rightarrow (1,2)} - 0.040 \cdot T'_{3,(1,2) \rightarrow (1,3)} + 0.002 \cdot T'_{3,(1,3) \rightarrow (2,3)} \\ & + 0.157 \cdot T'_{4,(1,0) \rightarrow (1,1)} + 0.156 \cdot T'_{4,(1,1) \rightarrow (1,2)} \\ & + 0.201 \cdot T'_{5,(1,0) \rightarrow (1,1)} - 0.010 \cdot T'_{5,(1,1) \rightarrow (0,1)} \end{aligned}$$

$$- 0.001 \cdot T'_{6,(1,3) \rightarrow (2,3)} + 0.000 \cdot T'_{6,(2,3) \rightarrow (3,3)} + 0.978.$$

The optimal integrated risk $R(\hat{\Theta}_y^* | y_{[N]})$ based on squared error is then 0.172 (which is the sum of the expected variance 0.097 and the expected squared bias 0.075).

2.2.2 Segment-Based and Route-Based Estimators

As we mentioned above, the optimal estimator $\hat{\Theta}_y^*$ uses historical traversal data on segments that are not part of route y . Moreover, the traversal data on different segments cannot be easily aggregated to obtain the optimal estimator, since the weight of each historical observation is non-trivially determined by the correlation structure. This makes the optimal estimator intractable to implement for large road networks, which typically consist of millions of road segments. Moreover, as we discussed in the previous subsection, the form of the optimal estimator depends on distributional assumptions, which can be hard to obtain in practice.

On the other hand, although being sub-optimal in general, the estimators mentioned in Example 1 are simple in nature and only use historical traversal data that is directly relevant to route y ; approaches like these are generally regarded as efficient and scalable methods and are practiced widely in mapping services. In addition, as we will show later, optimal estimators within these classes can be developed without distributional assumptions. It is thus of both practical and theoretical interests to compare their relative performance and benchmark them against the best estimator possible. To do so, generalizing the discussions in Example 1 and Corollary 1, we first introduce a formal definition of the segment-based estimators.

Definition 1 (SEGMENT-BASED ESTIMATOR). A segment-based estimator $\hat{\Theta}_y^{(\text{seg})}$ takes the form

$$\hat{\Theta}_y^{(\text{seg})} := \sum_{s \in y} \hat{\theta}_s, \quad \hat{\theta}_s := (1 - \phi_s(N_s))\mu + \phi_s(N_s) \frac{\sum_{n: s \in y_n} T'_{n,s}}{N_s},$$

for some $\phi_s : \mathbb{Z}_{\geq 0} \mapsto \mathbb{R}$ such that $\phi_s(0) = 0$ for all $s \in y$, and define $0/0 = 0$.

In other words, $\hat{\Theta}_y^{(\text{seg})}$ is the summation of segment-level estimators $\hat{\theta}_s$ that are constructed using a weighted average of the sample mean and the mean of the population distribution, where the weights $\{\phi_s(N_s)\}_{s \in y}$ are sample size dependent. One would typically expect the weights $\{\phi_s(N_s)\}_{s \in y} \in [0, 1]$ to converge to 1 as the sample size N_s grows to infinity. Such behavior ensures consistency of the estimator, i.e., $\hat{\theta}_s$ converges to θ_s almost surely as N_s tends to infinity. As an example, the estimator in Corollary 1 takes the form of $\phi_s(N_s) = \tau^2 N_s / (\tau^2 N_s + \sigma_s^2)$. Although the segment-based estimator in Definition 1 appears quite simple, variants on this estimator are used widely as a foundational component of commercial ETA prediction systems. In particular, either the time or speed is averaged for both (a) traversals of the road segment in the last few minutes; (b) traversals of the road segment from the same time of week in previous weeks. These are combined using either simple fallback logic (if sufficient recent traversals are available, then use their average, otherwise use the historical average) or a machine learning model trained with those features as inputs (see Section 3.1.3 and 4.1.2 of Derrow-Pinion et al. 2021).

Generalizing Definition 1 by combining individual segments into super-segments, we have the following definition of the generalized segment-based estimators. A super-segment S is a set of (usually connecting) distinct segments. Let \mathcal{S}_y denote a set of super-segments that constitutes route y , i.e., $\cup_{S \in \mathcal{S}_y} S = y$ and $S \cap T = \emptyset$ for any $S \neq T \in \mathcal{S}_y$. With a slight abuse of notation, let $N_S := |\{y_n : S \subset y_n, n \in [N]\}|$ be the sample size of traversals on super-segment S in the historical data.

Definition 2 (GENERALIZED SEGMENT-BASED ESTIMATOR). A generalized segment-based estimator $\hat{\Theta}_y^{(\text{g-seg})}$ takes the form

$$\hat{\Theta}_y^{(\text{g-seg})} := \sum_{S \in \mathcal{S}_y} \hat{\theta}_S, \quad \hat{\theta}_S := (1 - \phi_S(N_S))|S|\mu + \phi_S(N_S) \frac{\sum_{n: S \subset y_n} \sum_{s \in S} T'_{n,s}}{N_S},$$

for some \mathcal{S}_y , a set of super-segments constituting route y and some $\phi_S : \mathbb{Z}_{\geq 0} \mapsto \mathbb{R}$ such that $\phi_S(0) = 0$ for all $S \in \mathcal{S}_y$, and define $0/0 = 0$.

Similarly to the segment-based estimator, $\{\phi_S(N_S)\}_{S \in \mathcal{S}_y} \in [0, 1]$ are expected to converge to 1 as the sample size N_S approaches infinity. When $\mathcal{S}_y = \{\{s\}\}_{s \in y}$, the generalized segment-based estimator based on \mathcal{S}_y reduces to the segment-based estimator in Definition 1. A generalized segment-based travel time estimation method is described in Derrow-Pinion et al. [2021] for Google Maps.

We finally define a family of *route-based estimators* which uses route-level traversal data to estimate the travel time on a new route y . Let $\delta(y) \subset y_{[N]}$ denote a subset of historical routes which represents the neighborhood of route y . For example, $\delta(y)$ can be historical routes that share the same or similar origin and destination (but possibly with a different sequence of segments) as those of y (see Wang et al. 2016). These neighboring routes are representative observations to estimate travel time on route y . Let $M_{\delta(y)} = \sum_{n=1}^N \mathbf{1}\{y_n \in \delta(y)\}$ be the sample size of route y 's neighborhood, and $|y|$ be the number of segments traversed on route y .

Definition 3 (ROUTE-BASED ESTIMATOR). A route-based estimator $\hat{\Theta}_y^{(\text{route})}$ takes the form

$$\hat{\Theta}_y^{(\text{route})} := (1 - \phi_{\delta(y)}(M_{\delta(y)}))|y|\mu + \phi_{\delta(y)}(M_{\delta(y)}) \frac{\sum_{n: y_n \in \delta(y)} \sum_{s \in y_n} T'_{n,s}}{M_{\delta(y)}},$$

for some neighborhood of route y , $\delta(y)$, and some $\phi_{\delta(y)} : \mathbb{Z}_{\geq 0} \mapsto \mathbb{R}$ such that $\phi_{\delta(y)}(0) = 0$. Define $0/0 = 0$.

In words, $\hat{\Theta}_y^{(\text{route})}$ estimates travel time on route y by a weighted average of the sample mean of all observed travel times of the historical routes in $\delta(y)$, and the population mean of travel time on route y , where the weight $\phi_{\delta(y)}(M_{\delta(y)})$ is sample size dependent. Such a nearest-neighbor route-based estimator is used in Wang et al. [2016], for example.

We can derive the integrated risks $R(\hat{\Theta}_y^{(\text{g-seg})} | y_{[N]})$ and $R(\hat{\Theta}_y^{(\text{route})} | y_{[N]})$ conditional on historical routes $y_{[N]}$. With a slight abuse of notation, let $N_s^{\delta(y)} := |\{y_n \in \delta(y) : s \in y_n\}|$ be the number of traversals on segment s from the historical routes in neighborhood $\delta(y)$. Note that $N_s^{\delta(y)}$ is defined for $s \notin y$ as well since routes in $\delta(y)$ can traverse segments that are not in y . Similarly, for a set of distinct segments S , let $N_S^{\delta(y)} := |\{y_n \in \delta(y) : S \subset y_n\}|$ denote the number of traversals on super-segments S from the historical routes in neighborhood $\delta(y)$. By definition, we have $N_S^{\delta(y)} \leq N_s^{\delta(y)} \leq M_{\delta(y)}$, for all super-segments S such that $s \in S$. Let $\mathcal{S}_{\delta(y)} = \cup_{y' \in \delta(y)} y'$ be the set of all segments traversed by the routes in the neighborhood $\delta(y)$. Finally, put $\bar{y}_{\delta(y)} = \sum_{y_n \in \delta(y)} |y_n| / M_{\delta(y)}$ to be the average number of segments traversed per route in the neighborhood $\delta(y)$. We now give the integrated risks $R(\hat{\Theta}_y^{(\text{g-seg})} | y_{[N]})$ and $R(\hat{\Theta}_y^{(\text{route})} | y_{[N]})$ conditional on historical routes $y_{[N]}$. The integrated risk of the segment-based method $R(\hat{\Theta}_y^{(\text{seg})} | y_{[N]})$ is a special case of $R(\hat{\Theta}_y^{(\text{g-seg})} | y_{[N]})$ with $\mathcal{S}_y = \{\{s\}\}_{s \in y}$. It is worth noting that these expressions are derived under *no* distributional assumption.

Proposition 1. Under Assumptions 1 and 3, for any route y , the integrated risks, conditional on the historical routes $y_{[N]}$, are

$$R\left(\hat{\Theta}_y^{(\text{g-seg})} \mid y_{[N]}\right) = \sum_{S,T \in \mathcal{S}_y} \frac{N_{S \cup T}}{N_S N_T} \phi_S(N_S) \phi_T(N_T) \left(\sum_{s \in S, t \in T} \sigma_{s,t} \right) + \sum_{S \in \mathcal{S}_y} (1 - \phi_S(N_S))^2 |S| \tau^2, \quad (3)$$

$$R\left(\hat{\Theta}_y^{(\text{route})} \mid y_{[N]}\right) = \left(\frac{\phi_{\delta(y)}(M_{\delta(y)})}{M_{\delta(y)}} \right)^2 \left(\sum_{s,t \in \mathcal{S}_{\delta(y)}} N_s^{\delta(y)} \sigma_{s,t} \right) + (\phi_{\delta(y)}(M_{\delta(y)}) (\bar{y}_{\delta(y)} - |y|) \mu)^2 + \sum_{s \in \mathcal{S}_{\delta(y)} \setminus y} \left(\phi_{\delta(y)}(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2 + \sum_{s \in y} \left(1 - \phi_{\delta(y)}(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2. \quad (4)$$

The first and second terms in (3) correspond to the expected variance and squared bias of the generalized segment-based estimator, respectively. The expected squared bias comes from the shrinkage towards the prior mean $|S|\mu$ (can be different from the true means $\sum_{s \in S} \theta_s$), which goes down as $\phi_S(N_S)$ increases. The choice of $\phi_S(N_S)$ controls the bias-variance trade-off. Higher $\phi_S(N_S)$ (less shrinkage) leads to lower bias but introduces more variance as the estimator puts more weight on the information provided by the historical observations. Similarly, the first term in (4) represents the expected variance of the route-based estimator, and the sum of the second, third, and fourth terms collectively represents the expected squared bias of the route-based estimator. Specifically, the second term represents the squared bias introduced by including routes in $\delta(y)$ that have more or fewer road segments than y . The third term accounts for the squared bias of using traversal data on segments that are not included in y . Finally, the fourth term calculates the amount of squared bias induced by the shrinkage towards the prior mean $|y|\mu$. In addition to $\phi_{\delta(y)}(M_{\delta(y)})$, the choice of neighborhood $\delta(y)$ also plays a significant role here. If $\delta(y)$ is chosen to include only routes that are very similar to y , in terms of the number of segments and the set of segments they traverse, $\bar{y}_{\delta(y)}$ will be close to $|y|$, and $N_s^{\delta(y)}/M_{\delta(y)}$ will be close to 1 for segments $s \in y$ and close to 0 for segments $s \notin y$. This will lead to a lower bias, but potentially a higher variance as the number of samples $M_{\delta(y)}$ will be smaller.

Based on the formulae of the integrated risks, we define the optimal generalized segment-based estimator $\hat{\Theta}_y^{*(\text{g-seg})}$ given \mathcal{S}_y as the one that minimizes the integrated risk (3) by picking the best forms of $\{\phi_S(N_S)\}_{S \in \mathcal{S}_y}$. Also, given a neighborhood $\delta(y)$, the optimal route-based estimator $\hat{\Theta}_y^{*(\text{route})}$ is defined to be the one that minimizes the integrated risk (4) by picking the best form of $\phi_{\delta(y)}(M_{\delta(y)})$. The next result characterizes the forms of $\hat{\Theta}_y^{*(\text{g-seg})}$ and $\hat{\Theta}_y^{*(\text{route})}$. Note that $\hat{\Theta}_y^{*(\text{seg})}$ is a special case of $\hat{\Theta}_y^{*(\text{g-seg})}$ with $\mathcal{S}_y = \{\{s\}\}_{s \in y}$. Again, in contrast to Theorem 1, these optimal estimators are characterized under no distributional assumption.

Proposition 2. Under Assumptions 1 and 3, given \mathcal{S}_y , the optimal generalized segment-based estimator $\hat{\Theta}_y^{*(\text{g-seg})}$ takes the following form:

$$\hat{\Theta}_y^{*(\text{g-seg})} := \sum_{S \in \mathcal{S}_y} \hat{\theta}_S, \quad \hat{\theta}_S := (1 - \phi_S^*(N_S)) |S| \mu + \phi_S^*(N_S) \frac{\sum_{n: S \subset y_n} \sum_{s \in S} T'_{n,s}}{N_S}, \quad (5)$$

where $\{\phi_S^*(N_S)\}_{S \in \mathcal{S}_y}$ uniquely solves a system of linear equations

$$\sum_{T \in \mathcal{S}_y} (N_{S \cup T} / (N_S N_T)) \phi_T^*(N_T) \left(\sum_{s \in S, t \in T} \sigma_{s,t} \right) + (\phi_S^*(N_S) - 1) |S| \tau^2 = 0, \quad \forall S \in \mathcal{S}_y. \quad (6)$$

On the other hand, the optimal route-based estimator $\hat{\Theta}_y^{*(\text{route})}$ has the following form:

$$\hat{\Theta}_y^{*(\text{route})} := (1 - \phi_{\delta(y)}^*(M_{\delta(y)})) |y| \mu + \phi_{\delta(y)}^*(M_{\delta(y)}) \frac{\sum_{n: y_n = y} \sum_{s \in y} T'_{n,s}}{M_{\delta(y)}}, \quad (7)$$

$$\phi_{\delta(y)}^*(M_{\delta(y)}) = \left(\sum_{s \in y} N_s^{\delta(y)} \right) \tau^2 / \left(\sum_{s \in \mathcal{S}_{\delta(y)}} \frac{(N_s^{\delta(y)})^2}{M_{\delta(y)}} \tau^2 + \frac{\sum_{n: y_n \in \delta(y)} \sum_{s,t \in y_n} \sigma_{s,t}}{M_{\delta(y)}} + M_{\delta(y)} \mu^2 (\bar{y}_{\delta(y)} - |y|)^2 \right).$$

The following example illustrates the forms of these estimators.

Example 3. Using the same setup in Example 1 without the normality assumptions, we first compute the optimal segment-based estimator $\hat{\Theta}_y^{*(\text{seg})}$ for route y traversing through $(1, 0) \rightarrow (1, 1) \rightarrow (1, 2)$. Note that the covariance matrix $e^{-\mathcal{L}}$ has all non-negative entries. According to Proposition 2 of the supplement,

$$\hat{\Theta}_y^{*(\text{seg})} = \left(0.560 \cdot \frac{T'_{1,s_1} + T'_{4,s_1} + T'_{5,s_1}}{3} + 0.440 \cdot \mu \right) + \left(0.562 \cdot \frac{T'_{2,s_2} + T'_{3,s_2} + T'_{4,s_2}}{3} + 0.438 \cdot \mu \right),$$

with integrated risk 0.176 (expected variance 0.099 and expected squared bias 0.077).

The optimal generalized segment-based estimator $\hat{\Theta}_y^{*(\text{g-seg})}$ based on $\mathcal{S}_y = \{\{y\}\}$ is with $\phi_y^*(N_y) = N_y |y| \tau^2 / (N_y |y| \tau^2 + \sum_{s,t \in y} \sigma_{s,t})$. In this example, $N_y = 1$. This gives

$$\hat{\Theta}_y^{*(\text{g-seg})} = 0.267 \cdot (T'_{4,s_1} + T'_{4,s_2}) + 0.733 \cdot (2\mu),$$

with integrated risk 0.293 (expected variance 0.078 and expected squared bias 0.215).

Finally, consider the optimal route-based estimator $\hat{\Theta}_y^{*(\text{route})}$ based on $\delta(y)$ which includes all historical routes whose origin and destination are at almost one segment away from that of y . This includes trips \mathcal{T}_4 and \mathcal{T}_5 .

$$\hat{\Theta}_y^{*(\text{route})} = 0.372 \cdot \frac{(\sum_{s \in y} T'_{4,s} + \sum_{s \in y} T'_{5,s})}{2} + 0.628 \cdot (2\mu),$$

with integrated risk 0.288 (expected variance 0.070 and expected squared bias 0.218).

Our next result investigates the comparison of the integrated risks of these estimators, under a case where $\sigma_{s,t} \geq 0$ for all $s \neq t \in \mathcal{S}$, i.e., there exists only non-negative covariances in the road network. We first show that the optimal segment-based estimator $\hat{\Theta}_y^{*(\text{seg})}$ is more accurate than a wide variety of optimal route-based estimators $\hat{\Theta}_y^{*(\text{route})}$.

Theorem 2. *In addition to Assumptions 1 and 3, suppose that $\sigma_{s,t} \geq 0$ for all $s, t \in \mathcal{S}$. Let $\hat{\Theta}_y^{*(\text{seg})}$ be the optimal segment-based estimator and let $\hat{\Theta}_y^{*(\text{route})}$ be the optimal route-based estimator with neighborhood $\delta(y)$ such that*

$$N_{s \cup t} N_s^{\delta(y)} N_t^{\delta(y)} \leq N_{s \cup t}^{\delta(y)} N_s N_t, \quad \forall s, t \in y, \quad (8)$$

we have,

$$R\left(\hat{\Theta}_y^{*(\text{seg})} \mid y_{[N]}\right) \leq R\left(\hat{\Theta}_y^{*(\text{route})} \mid y_{[N]}\right),$$

for any set of historical routes $y_{[N]}$.

Condition (8) on the neighborhood of route-based estimators $\delta(y)$ is very mild — by re-arranging the term, we have

$$N_{s \cup t} / (N_s N_t) \leq N_{s \cup t}^{\delta(y)} / (N_s^{\delta(y)} N_t^{\delta(y)}), \quad \forall s, t \in y.$$

One can effectively think of $N_{s \cup t}$, N_s , and N_t as the sample sizes under a neighborhood that includes *all* historical routes. In other words, condition (8) intuitively requires the chosen neighborhood $\delta(y)$ to concentrate more around the predicting route compared to a neighborhood which includes all routes — the ratio $N_{s \cup t}^{\delta(y)} / (N_s^{\delta(y)} N_t^{\delta(y)})$ gets larger as the neighborhood $\delta(y)$ concentrates around y . This is generally expected because as the neighborhood gets smaller, routes in $\delta(y)$ become more similar to y . It then becomes more likely that a route in $\delta(y)$ traversing over segment $s \in y$ also traverses over segment $t \in y$. Another way to appreciate this intuition is to look at the other extreme — the smallest possible neighborhood $\delta(y) = \{y_n : y_n = y\}$ which only includes historical routes that are *exactly the same* as route y . Under such a neighborhood, we have $N_{s \cup t}^{\delta(y)} = N_s^{\delta(y)} = N_t^{\delta(y)}$. Now consider any neighborhood $\delta'(y) \supset \delta(y)$, we have

$$\frac{N_{s \cup t}^{\delta'(y)}}{N_s^{\delta'(y)} N_t^{\delta'(y)}} \leq \frac{N_{s \cup t}^{\delta(y)}}{N_s^{\delta(y)} N_t^{\delta(y)}} \Leftrightarrow N_{s \cup t}^{\delta'(y)} N_t^{\delta(y)} \leq N_s^{\delta'(y)} N_t^{\delta'(y)}, \quad \forall s, t \in y.$$

The latter holds because $N_{s \cup t}^{\delta'(y)} \leq N_s^{\delta'(y)}$ and $N_t^{\delta(y)} \leq N_t^{\delta'(y)}$, $\forall s, t \in y$. In other words, enlarging the neighborhood from the smallest one $\delta(y) = \{y_n : y_n = y\}$ always decreases the ratio $N_{s \cup t}^{\delta(y)} / (N_s^{\delta(y)} N_t^{\delta(y)})$.

We now show that the non-negative covariance assumption in Theorem 2 is critical. When travel times on different road segments can potentially be negatively correlated, we show, with the following example, that the optimal segment-based estimator can produce a strictly higher integrated risk than the optimal route-based estimator with neighborhood $\delta(y) = \{y_n : y_n = y\}$.

Example 4. Using the same setup in Example 1 without normality assumptions, we now compare the integrated risk of $\hat{\Theta}_y^{*(\text{seg})}$ with that of $\hat{\Theta}_y^{*(\text{route})}$ with a neighborhood $\delta(y) = \{y_n : y_n = y\}$, under negative covariances. Suppose $\sigma_{s_1}^2 = \sigma_{s_2}^2 = 1$ and $\sigma_{s_1, s_2} = \sigma_{s_2, s_1} = -0.9$. Let $\tau^2 = 1$. In this case, the optimal segment-based estimator $\hat{\Theta}_y^{*(\text{seg})}$ takes the form

$$\begin{aligned} \hat{\Theta}_y^{*(\text{seg})} = & 0.811 \cdot \frac{T'_{1, s_1} + T'_{4, s_1} + T'_{5, s_1}}{3} + 0.189 \cdot \mu \\ & + 0.811 \cdot \frac{T'_{2, s_2} + T'_{3, s_2} + T'_{4, s_2}}{3} + 0.189 \cdot \mu, \end{aligned}$$

with integrated risk 0.378 (expected variance 0.307 and expected squared bias 0.071).

On the other hand, the optimal route-based estimator $\hat{\Theta}_y^{*(\text{route})}$ takes the form,

$$\hat{\Theta}_y^{*(\text{route})} = 0.909 \cdot (T'_{4, s_1} + T'_{4, s_2}) + 0.091 \cdot (2\mu),$$

with integrated risk 0.182 (expected variance 0.165 and expected squared bias 0.017).

The intuition behind the observation that negatively correlated segment travel time can benefit the route-based estimator is that route-level travel times can potentially *absorb* the variance of segment travel times by avoiding additional aggregation. This could sometimes create an edge over the segment-based estimator even when the route-based estimator uses fewer samples. Negative correlations between the segments can occur, for instance, due to having traffic signals in the route. If one segment is slow due to a red signal, the subsequent segment can have faster travel time due to a green signal [Ramezani and Geroliminis, 2012].

Based on this observation in Example 4, it is reasonable to conjecture that when all the covariances are non-negative $\sigma_{s,t} \geq 0, \forall s, t \in \mathcal{S}$, the optimal segment-based estimator $\hat{\Theta}_y^{*(\text{seg})}$ has the minimum integrated risk among *all* generalized segment-based estimators $\hat{\Theta}_y^{(\text{g-seg})}$. It appears at first glance that aggregating segment travel times into super-segment travel times in this case does not help reduce the overall variance of the estimator. Surprisingly, the next example shows that this might *not* be the case.

Example 5. Using the same setup in Example 1 without normality assumptions, we consider the optimal segment-based and generalized segment-based estimator for the travel time of a new route y traversing through $(1, 2) \rightarrow (1, 3) \rightarrow (2, 3) \rightarrow (3, 3)$. We call segment $(1, 2) \rightarrow (1, 3)$ to be s_3 , segment $(1, 3) \rightarrow (2, 3)$ to be s_4 and segment $(2, 3) \rightarrow (3, 3)$ to be s_5 . Consider $\sigma_{s_3}^2 = 0.1, \sigma_{s_4}^2 = 10, \sigma_{s_5}^2 = 10, \sigma_{s_3, s_4} = 1$, and $\sigma_{s_3, s_5} = \sigma_{s_4, s_5} = 0$. Let $\tau^2 = 1$. One can check that this is a valid covariance matrix. We first compute the optimal segment-based estimator $\hat{\Theta}_y^{*(\text{seg})}$, which takes the form

$$\hat{\Theta}_y^{*(\text{seg})} = (0.866 \cdot T'_{3, s_3} + 0.134 \cdot \mu) + \left(0.094 \cdot \frac{T'_{3, s_4} + T'_{6, s_4}}{2} + 0.906 \cdot \mu \right) + (0.091 \cdot T'_{6, s_5} + 0.909 \cdot \mu).$$

The integrated risk of $\hat{\Theta}_y^{*(\text{seg})}$ is 1.948 (expected variance 0.284 and expected squared bias 1.664).

Now the optimal generalized segment-based estimator $\hat{\Theta}_y^{*(\text{g-seg})}$ with $\mathcal{S}_y = \{\{s_3\}, \{s_4, s_5\}\}$. It takes the form

$$\hat{\Theta}_y^{*(\text{g-seg})} = (0.909 \cdot T'_{3, s_3} + 0.091 \cdot \mu) + (0.091 \cdot (T'_{6, s_4} + T'_{6, s_5}) + 0.909 \cdot (2\mu)). \quad (9)$$

The integrated risk of $\hat{\Theta}_y^{*(\text{g-seg})}$ is 1.909 (expected variance 0.248 and expected squared bias 1.661).

The reason that the optimal segment-based estimator is not the best among all generalized segment-based estimators under non-negative covariances is quite subtle. In Example 5, the only pair of segments that are correlated are s_3 and s_4 with a positive covariance 1. Interestingly, merging s_4 and s_5 into a super-segment avoids increasing the variance of the estimator resulting from the positive covariance between a *different* pair of segments s_3 and s_4 . To see that, in the form of $\hat{\Theta}_y^{*(\text{g-seg})}$ (equation (9)), historical traversal data on segments s_3 and s_4 from trip \mathcal{T}_3 are not both used in the estimator because \mathcal{T}_3 does not traverse through all three segments s_3, s_4 and s_5 . In other words, merging segments into super-segments sometimes breaks the dependency of two segments within a different super-segment. This is achieved by creating a higher barrier for the historical traversals on these segments within the same trip to be included in the estimator.

Nevertheless, we have the following proposition when the optimal generalized segment-based estimator uses the entire route y as a super-segment, i.e., $\mathcal{S}_y = \{\{y\}\}$, and the optimal route-based estimator $\hat{\Theta}_y^{*(\text{route})}$ uses the neighborhood $\delta(y) = \{y_n : y_n = y\}$ that contains the exact same route as y in the historical data. There is a (subtle) difference between these two estimators. The former includes all traversals that go through y , i.e., y is a sub-path of the traversals. On the other hand, the latter only includes historical routes that share the exact same route as y including its origin and destination.

Proposition 3. In addition to Assumptions 1 and 3, suppose that $\sigma_{s,t} \geq 0$ for all $s, t \in y$. Let $\hat{\Theta}_y^{*(\text{seg})}$ be the optimal segment-based estimator, $\hat{\Theta}_y^{*(\text{g-seg})}$ be the optimal generalized segment-based estimator with $\mathcal{S}_y = \{\{y\}\}$, and $\hat{\Theta}_y^{*(\text{route})}$ be the optimal route-based estimator with neighborhood $\delta(y) = \{y_n : y_n = y\}$,

$$R\left(\hat{\Theta}_y^{*(\text{seg})} \middle| y_{[N]}\right) \leq R\left(\hat{\Theta}_y^{*(\text{g-seg})} \middle| y_{[N]}\right) \leq R\left(\hat{\Theta}_y^{*(\text{route})} \middle| y_{[N]}\right),$$

for any set of historical routes $y_{[N]}$.

We conclude this section by commenting that although Theorem 2 and Proposition 3 give some evidence in terms of the superiority of the optimal segment-based estimator, Example 4 and Example 5 also point out that there are cases where the comparisons are not clean. To garner more insights, in the next section, we are going to analyze an *asymptotic* setting where the number of trip observations grows with the size of the road network.

3 Asymptotic Analysis

In this section, we compare estimators in terms of how their integrated risks scale with the road network size. We consider an asymptotic setting where the number of trip observations grows with the size of the road network.² This regime is relevant in practice, since the road network of a major metropolitan area typically contains hundreds of thousands to millions of road segments, and typical commercial datasets contain tens of millions of trips in such a network [Li et al., 2018]. One benefit of such an asymptotic analysis is to compare estimators in a more tractable setting, enabling comparisons that can't be done in the finite-sample setting.

We start by pointing out that the optimal segment-based estimator $\hat{\Theta}_y^{*(\text{seg})}$ requires inverting a $|y| \times |y|$ matrix which could be computationally intensive for real-time implementation on large-scale road networks. Moreover, it also requires explicit knowledge of the covariance structures among each pair of road segments $\sigma_{s,t}$, which can be hard to precisely estimate in practice. Our goal in this section is to see if a similar (or stronger) result as Theorem 2 or Proposition 3 holds in an asymptotic limit with a class of *much simpler* segment-based estimators. These simple segment-based estimators are tractable to compute for large road networks and do not require any knowledge of the covariance structures. In addition, we aim to generalize our result to a case where there exist *negative* correlations among segment travel times. Finally, the asymptotic analysis also enables us to compare the segment-based estimators with generalized segment-based estimators, which we are not able to do in the finite-sample case.

We first introduce our asymptotic setting. Consider a road network indexed by a size $p \in \mathbb{N}$, with a set of vertices (intersections) \mathcal{V}_p and a set of edges (road segments) \mathcal{S}_p . An example road network is the grid network where p represents the size of the grid. For a grid network with size p , $|\mathcal{S}_p| = |\mathcal{V}_p| \simeq p^2$. Let \mathcal{Y}_p be the set of all possible routes in the road network of size p . We assume that any route $y \in \mathcal{Y}_p$ contains at least one road segment, $|y| \geq 1$. The number of trips N in the training data grows with p , such that $N \rightarrow \infty$ as $p \rightarrow \infty$, although this is not strictly required for any of the following results. In addition,

Assumption 4. Assume that:

²One can also consider the context of a fixed size road network, and analyze the efficiency as the number of trips $N \rightarrow \infty$. This is less informative because nearly all reasonable approaches have the same asymptotic rate as a function of N , but with very large (and meaningful) differences in their constants.

1. For each road network with size p , the historical routes $Y_{p,[N]}$ in the training data as well as the predicting route Y_p are drawn independently according to some probability distribution μ_p over \mathcal{Y}_p .
2. The covariance matrix $\Sigma_p = [\sigma_{s,t}]_{s,t \in \mathcal{S}_p}$ and its corresponding precision matrix $\Psi_p = \Sigma_p^{-1} = [\psi_{s,t}]_{s,t \in \mathcal{S}_p}$ under network size p satisfy $\sum_{t \in \mathcal{S}_p} |\sigma_{s,t}| = \mathcal{O}(1)$ and $\sum_{t \in \mathcal{S}_p} |\psi_{s,t}| = \mathcal{O}(1)$, $\forall s \in \mathcal{S}_p$. Moreover, there exists $\sigma_{\min} > 0$ such that for any route $y_p \in \mathcal{Y}_p$, $\sum_{s,t \in y_p} \sigma_{s,t} \geq \sigma_{\min}$.

The first part of the assumption introduces a route distribution μ_p for each size of the road network, from which historical routes and predicting routes are sampled. Note that $Y_{p,[N]}$ and Y_p are capitalized because they are random in this setting. The second part of the assumption is justifiable in the ETA prediction context since spatial decay in the correlation of segment travel times is widely observed in empirical studies — the correlation between two road segments decays as the distance between the two segments increases (see e.g., Bernard et al. 2006, Rachtan et al. 2013, Guo et al. 2020, Woodard et al. 2017). It further implies that the sum of all the (co)variance components in the road network grows at most linearly to the total number of segments, $\sum_{s,t \in \mathcal{S}_p} \sigma_{s,t} \leq \sum_{s,t \in \mathcal{S}_p} |\sigma_{s,t}| = \mathcal{O}(|\mathcal{S}_p|)$.

Similarly to Section 2, we compare the accuracy of different estimators using integrated risk. To obtain our results in this asymptotic setting where routes are randomly sampled, we slightly alter the definition of the integrated risk used in Section 2. Specifically, for a given road network of size p , we leverage Assumption 4 to integrate the risk over the distribution of historical routes $Y_{p,[N]}$ and predicting route Y_p . This yields the following definition of integrated risk:

$$R(\hat{\Theta}_{Y_p}) := \mathbb{E} \left[\left(\hat{\Theta}_{Y_p} - \Theta_{Y_p} \right)^2 \right], \quad (10)$$

where the expectation is now taken with respect to (1) the distribution over the historical routes $Y_{p,[N]}$, (2) the predicting route Y_p , in addition to (3) the adjusted travel times $\{T'_{n,s}\}_{n \in [N], s \in Y_{p,n}}$ and (4) the population distribution on $\{\theta_s\}_{s \in \mathcal{S}_p}$. Our results will compare the asymptotic integrated risk of travel time estimators $R(\hat{\Theta}_{Y_p})$ as $p \rightarrow \infty$ (and $N \rightarrow \infty$).

3.1 Grid Networks

We consider an example of grid road networks. Let $x = (i, j) \in \mathcal{V}_p$ for $\mathcal{V}_p = \{0, \dots, p\}^2$ denote a vertex on the grid (a possible start or end point of a route), and $s \in \mathcal{S}_p$ denote a road segment, i.e., a directed edge between adjacent vertices. We define the route distribution μ_p under grid size p by assuming that the trip's origin $x_1 = (i_1, j_1)$ and destination $x_2 = (i_2, j_2)$ are drawn independently from the following probability distribution over vertices:

$$\mathbb{P}[X = (i, j)] = \prod_{k \in \{i, j\}} \binom{p}{k} \frac{B(\alpha + k, \alpha + p - k)}{B(\alpha, \alpha)},$$

where $0 < \alpha \leq 1$ and $B(\cdot, \cdot)$ denotes the beta function.³ In other words, the east-west and north-south coordinates of the origin and destination are independently sampled from a *symmetric beta-binomial distribution*. When $\alpha < 1$, this distribution has a “horseshoe” shape, with a high probability at the edges of the grid and a low probability in the center. For $\alpha = 1$, this is just the

³The case of $\alpha > 1$ is less interesting as origins and destinations concentrate within the center of the grid, and so trips do not fully utilize the entire p by p grid. This case can somewhat be captured by a grid with a smaller size. Nevertheless, we look at the case of $\alpha > 1$ in the numerical experiments in Section 3.2.

uniform distribution over \mathcal{V}_p . As α decreases, the distribution more heavily weighs the locations near the four corners of the grid. Given the origin and the destination, routes are sampled from some distributions we do not put restrictions on first.

We consider a neighborhood $\delta^{\text{od}}(\cdot)$ that includes all historical routes whose origins and destinations are close to those of the predicting route respectively. We define $x_1(y_p)$ and $x_2(y_p)$ as the origin and destination of route y_p . Construct $\delta^{\text{od}}(y_p) = \{y \in \mathcal{Y}_p : \|x_1(y), x_1(y_p)\|_1 \leq c, \|x_2(y), x_2(y_p)\|_1 \leq c\}$ for some fixed constant $c > 0$ that does *not* depend on p . In our first asymptotic result below, we compare a large family of simple segment-based estimators $\hat{\Theta}_y^{(\text{seg})}$ to the optimal route-based estimators $\hat{\Theta}_y^{*(\text{route})}$ with neighborhood $\delta^{\text{od}}(\cdot)$. This family of simple segment-based estimators only requires that $\phi_s(N_s)$ approaches 1 quickly enough. We provide this result without restricting to non-negative covariances, as required in Theorem 2.

Theorem 3. *Under Assumptions 1, 3 and 4, consider an optimal route-based estimator $\hat{\Theta}_y^{*(\text{route})}$ based on a route neighborhood $\delta^{\text{od}}(\cdot)$ with similar origin and destination as the those of the predicting route, if $1/4 < \alpha \leq 1$,*

$$\lim_{p \rightarrow \infty} \frac{R\left(\hat{\Theta}_{Y_p}^{(\text{seg})}\right)}{R\left(\hat{\Theta}_{Y_p}^{*(\text{route})}\right)} = 0,$$

for any segment-based estimator $\hat{\Theta}_y^{(\text{seg})}$ with $\phi_s(N_s) = 1 - \mathcal{O}(1/\sqrt{N_s})$, $\forall s \in \mathcal{S}_p$. In addition,

$$R(\hat{\Theta}_{Y_p}^{(\text{seg})}) = \mathcal{O}(p^2/N), \quad R(\hat{\Theta}_{Y_p}^{*(\text{route})}) = \begin{cases} \Omega(p^4/N), & 1/2 < \alpha \leq 1, \\ \Omega(p^{8\alpha}/N), & 0 < \alpha \leq 1/2. \end{cases}$$

Remark 1. This result holds under *any* route distribution given origin and destination. In fact, the integrated risk of the simple segment-based estimator $R(\hat{\Theta}_{Y_p}^{(\text{seg})}) = \mathcal{O}(p^2/N)$ holds under *any* distribution of origins and destinations. This result also does *not* require any minimum data growth rate on N as a function of p , which suggests that the result holds under data-sparse settings.

Remark 2. The family of segment-based estimators considered in Theorem 3 includes, for example, the optimal estimator under independent, Gaussian distributed segment travel times in Corollary 1, $\phi_s(N_s) = N_s\tau^2/(N_s\tau^2 + \sigma_s^2)$. It is interesting to note that the rate requirement $\phi_s(N_s) = 1 - \mathcal{O}(1/\sqrt{N_s})$ gives some leeway in the sense that $\phi_s(N_s)$ can approach 1 more slowly than what is required in the optimal estimator for the independent case. This family of segment-based estimators also includes other simple forms without any knowledge of the variance parameters, e.g., $\phi_s(N_s) = N_s/(N_s + \lambda)$, for any $\lambda > 0$; or some threshold-based structure such as $\phi_s(N_s) = 1$ if $N_s \geq c$ with some constant $c > 0$ and $\phi_s(N_s) = 0$ otherwise. The former choice of $\phi_s(N_s)$ can be interpreted as an estimator $\hat{\Theta}_y$ minimizing penalized squared-error loss: $\min_{\hat{\theta}_s} \sum_{n:s \in y_n} (\hat{\theta}_s - T'_{n,s})^2 + \lambda(\hat{\theta}_s - \mu)^2$ where $\lambda > 0$ is the regularizing parameter; and the latter choice of $\phi_s(N_s)$ can be thought of as a simple fallback logic — predict the segment travel time using the sample average if sample size exceeds a threshold or using the population mean μ if there is not enough data. The simpler forms of $\phi_s(N_s)$ enhance the relevance of this result since, in practice, mapping services do not have access to the “optimal” form $\phi_s^*(N_s)$ and the choice of $\phi_s(N_s)$ is often tuned through cross-validation.

Remark 3. One can also develop a result similar to Theorem 3 where the size of the neighborhood $\delta^{\text{od}}(y_p) = \{y \in \mathcal{Y}_p : \|x_1(y), x_1(y_p)\|_1 \leq c, \|x_2(y), x_2(y_p)\|_1 \leq c\}$, c , grows with the grid size p . One approach is to simply multiply the sample size by the number of distinct origin-destination pairs in the neighborhood and revise Theorem 3.1 accordingly. However, this is an overly optimistic

lower bound of the integrated risk of the route-based method, because increasing the size of the neighborhood also introduces additional biases by including relatively irrelevant historical trips. In Section 3.2, we conduct numerical experiments to investigate this trade-off. These experiments show that having an increasing neighborhood size does not change the asymptotic comparisons in Theorem 3. This suggests that the additional biases introduced by a larger neighborhood can offset the benefits of a larger sample size.

Theorem 3 suggests that when the distributions of the route origins and destinations are not overly concentrated, a route-based estimator using routes with similar origins and destinations is *asymptotically dominated* by a class of simple segment-based estimators. When origins and destinations of the routes are too concentrated, historical and predicting can be very similar — in the extreme case where $\alpha \rightarrow 0^+$, all origins and destinations are concentrated at the four corners of the grid so that there are only 16 types of origin-destination pairs in the data where each route goes from one corner of the grid to another. This can give a route-based estimator some advantages. The requirement $1/4 < \alpha \leq 1$ ensures that there is enough dispersion among historical routes. This range is quite generous — when $\alpha = 1/4$, for a 10×10 grid, the probability of sampling route origins or destinations at the corners is over 50 times higher than the center of the grid, and this gap increases as the grid size increases.

It turns out that we are able to say a lot more by directly comparing the segment-based estimators $\hat{\Theta}_{Y_p}^{(\text{seg})}$ to the optimal estimator $\hat{\Theta}_{Y_p}^*$ characterized in Theorem 1. To do so, we first fully specify the route distribution μ_p . Conditional on the origin and destination $x_1 = (i_1, j_1)$ and $x_2 = (i_2, j_2)$, we sample the route Y_p uniformly from the set of all routes in \mathcal{Y}_p that *minimize both the number of traversals and turns* from x_1 to x_2 , i.e., that have length equal to the grid distance $\|x_1 - x_2\|_1 = |i_1 - i_2| + |j_1 - j_2|$ and the minimum number of turns. Figure 2 below illustrates the route distribution given specific origin x_1 and destination x_2 . On the left of Figure 2, there is only one possible route between them, while on the right of Figure 2, there are two possible routes, each with probability 0.5 being sampled. Although somewhat simplified, this route distribution μ_p biases towards route-based (and generalized segment-based) estimators as it significantly limits the set of possible routes \mathcal{Y}_p and increases the sample size of each possible route $y \in \mathcal{Y}_p$. In other words, for segment-based estimators, having good relative performance under such a route distribution μ_p likely implies good relative performance under other route distributions.

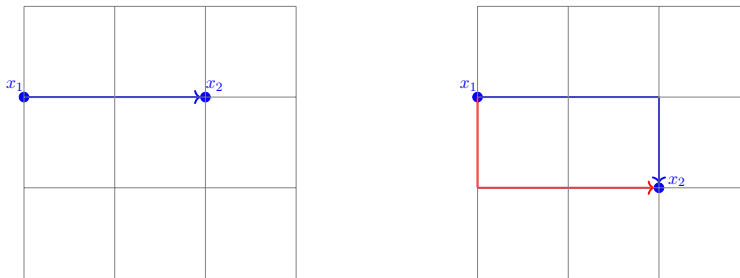


Figure 2: Examples of the route distribution μ_p conditional on origin x_1 and destination x_2 .

We now give the main result of the paper. We show that the same class of simple segment-based estimators considered in Theorem 3 is *asymptotically optimal up to a logarithmic factor*. Precisely, we compare its integrated risk with that of the optimal estimator $\hat{\Theta}_y^*$ under normal distributions and show that, in the asymptotic limit, the ratio of the risks can be upper bounded by a small logarithmic factor $\mathcal{O}(\log(p))$. Such comparison is non-trivial as we neither have a closed-form formula for the optimal estimator $\hat{\Theta}_{Y_p}^*$, nor for its risk (see Theorem 1). We thus compare $R(\hat{\Theta}_{Y_p}^{(\text{seg})})$

to a more tractable lower bound of the optimal risk $R(\hat{\Theta}_Y^*)$. This lower bound is obtained by adapting the Bayesian Cramér-Rao bound [Gill and Levit, 1995] through the van Trees inequality [van Trees, 2004], a Bayesian analog of the information inequality (see Lemma 1 in the supplement).

We now formally present the result regarding the asymptotic optimality of the segment-based estimators below.

Theorem 4 (ASYMPTOTIC OPTIMALITY OF SEGMENT-BASED ESTIMATORS). *In addition to Assumptions 1, 3 and 4, assume that (\mathcal{E}, θ) are jointly Gaussian distributed. When $1/2 \leq \alpha \leq 1$ and $N = \omega(p)$,*

$$\frac{R\left(\hat{\Theta}_{Y_p}^{(\text{seg})}\right)}{R\left(\hat{\Theta}_{Y_p}^*\right)} = \mathcal{O}(\log(p)),$$

for any segment-based estimator with $\phi_s(N_s) = 1 - \mathcal{O}(1/\sqrt{N_s})$, $\forall s \in \mathcal{S}_p$. When the data growth rate $N = \mathcal{O}(p)$, $\liminf_{p \rightarrow \infty} R\left(\hat{\Theta}_Y\right) > 0$ for any estimator $\hat{\Theta}_Y$.

Theorem 4 has strong practical implications. It says that although improvement can be made in some cases over simple segment-based estimators by, for example, using a route-based method (Example 4) or combining segments into super-segments (Example 5), their benefits are limited and in the asymptotic limit where grid size and sample size grow, these benefits can only make a difference up to a logarithmic factor. This gives reassurance that maintaining a segment-based travel time prediction architecture achieves most of the accuracy of the optimal estimator. Similar to Theorem 3, Theorem 4 does require some conditions to make sure that historical routes are diverse enough. The condition $1/2 \leq \alpha \leq 1$ is stricter than the one required in Theorem 3 but still quite generous — when $\alpha = 1/2$, for a 10×10 grid, the probability of sampling route origins or destinations at the corners is more than 8 times higher than in the center of the grid and this gap again increases as the grid size increases. In addition, Theorem 4 also requires the data growth rate to be at least $N = \omega(p)$. This turns out to be a very mild condition as for any slower data growth rate $N = \mathcal{O}(p)$, no estimator can be consistent in the sense that the asymptotic risk tends to zero.

3.2 Numerical Examples

We numerically demonstrate the accuracy of different estimators based on representative correlation structures used for travel times on road networks. We construct $p \times p$ grid networks where $p \in \{10, 15, 20, 25, 30\}$. For each grid size p , we consider different sample sizes of historical routes $N = p, p^2, p^3$, and p^4 . Each historical route is generated from a route distribution μ_p as detailed at the beginning of Section 3.1. We consider a route distribution with $\alpha = 1.0$ — origins and destinations are generated uniformly over the grid. The covariance matrix of segment travel times is taken to be $ue^{-v\mathcal{L}} + I$ where $u, v > 0$ are some parameters and I is an identity matrix representing a white noise of travel time uncertainty. The matrix $\mathcal{L} = D^{-1/2}LD^{1/2}$ is the normalized Laplacian of the grid network where D is the diagonal matrix of segment degrees, A is the adjacency matrix of the grid network and $L = D - A$ is the graph Laplacian. This is also called the diffusion kernel. It models spatial decay of correlation among segment travel times. As u increases, the matrix becomes more diffused in the sense that the correlation becomes relatively stronger. On the other hand, v controls the weight between the diffusion kernel and the white noise. Note that this covariance structure also satisfies the second part of Assumption 4. We set the population variance of the means of segment travel times to be $\tau^2 = 0.5$, which is similar to the variances of the segment travel times σ_s^2 .

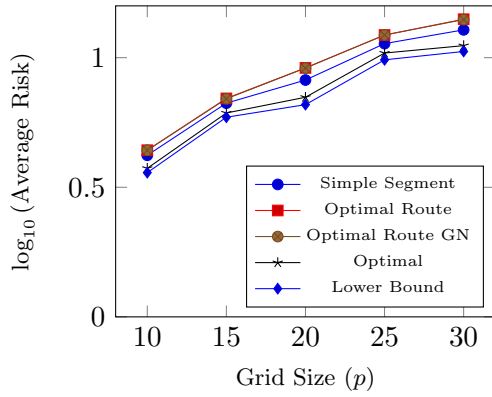
For each grid size $p \in \{10, 15, 20, 25, 30\}$ and taking the parameter of the route distribution to be $\alpha \in \{0.3, 1.0, 3.0\}$, we generate 100 predicting routes and report the average integrated risk of these predicting routes for the following methods, under a covariance matrix $ue^{-v\mathcal{L}} + I$ specified with $(u, v) = (1, 1)$.

1. Simple segment-based method $\hat{\Theta}_{y_p}^{(\text{simple-seg})}$ with $\phi_s(N_s) = N_s/(N_s + 1)$.
2. Optimal route-based method $\hat{\Theta}_{y_p}^{*(\text{route})}$ with $\delta(y_p) = \{y_n : x_1(y_n) = x_1(y_p), x_2(y_n) = x_2(y_p)\}$ that includes all historical routes sharing the same origin and destination with the predicting route y_p .
3. Optimal route-based method $\hat{\Theta}_{y_p}^{*(\text{route})}$ with a growing neighborhood $\delta(y_p) = \{y_n : \|x_1(y_n) - x_1(y_p)\|_1 \leq [0.1p], \|x_2(y_n) - x_2(y_p)\|_1 \leq [0.1p]\}$ that includes all historical routes sharing similar origin and destination with the predicting route y_p where the degree of similarity is growing with the grid size p .
4. Optimal estimator $\hat{\Theta}_{y_p}^*$ in Theorem 1 under the assumption that (\mathcal{E}, θ) are jointly Gaussian distributed.
5. The information-theoretic lower bound developed in Lemma 1 in the supplement under the assumption that (\mathcal{E}, θ) are jointly Gaussian distributed.

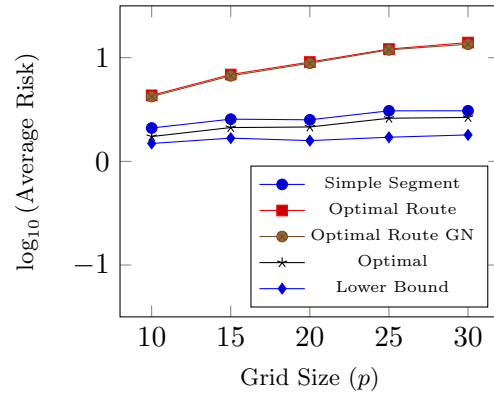
We reiterate that the average integrated risks of the simple segment-based method and the optimal route-based method do not depend on distributional assumptions. Figures 3, 4 and 5 report the average integrated risks (in logarithmic scale) of the aforementioned methods over 100 predicting routes, under different sample sizes as the grid size p increases. Each figure corresponds to a different route distribution. Figure 3 depicts the situation under $\alpha = 0.3$ where origins and destinations of the routes are more concentrated at the corners of the grid; Figure 4 represents $\alpha = 1.0$ where route origins and destinations are uniformly distributed over the grid; and finally Figure 5 reports the situation of $\alpha = 3.0$ where origins and destinations are more concentrated at the central part of the grid, though this is outside the range of $\alpha \in (0, 1]$ we assume in Section 3.1. These numerical findings match our theoretical results. The increasing difference between the integrated risks of the two optimal route-based methods and those of the simple segment-based method reflects the asymptotic dominance result in Theorem 3 as well as Remark 3 that the dominance results likely remain true even if we consider a route-based method with growing neighborhood size (marked by ‘‘Optimal Route GN’’). The average risks of the simple segment-based method tend to zero when the sample size grows faster than $N = p^{2.0}$. The performance of the simple segment-based estimator is extremely competitive — it almost matches the optimal risk. The gap between the risk of the simple segment-based method and the information-theoretic lower bound increases very mildly as grid size p increases, which reflects the logarithmic scaling in Theorem 4. It is worth noting that both the cases of $\alpha = 3.0$ and $\alpha = 0.3$ are outside the range of α assumed in Theorem 4 but the optimality seems to remain valid. In Appendix D of the supplement, we report additional numerical experiments under other covariance matrices $ue^{-v\mathcal{L}} + I$ specified with other values of (u, v) as well as covariance structures that violate the second part of Assumption 4.

4 Concluding Remarks

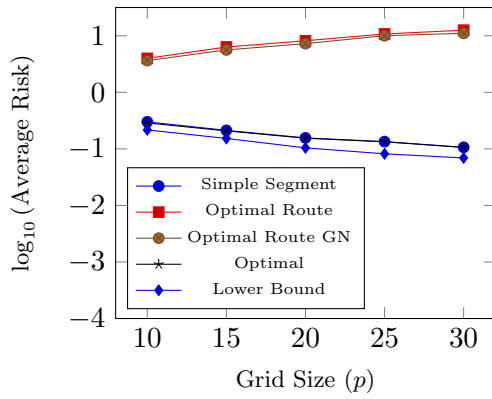
Our model and analysis reveal insights into the accuracy of various travel time predictors used in practice. Our results favor segment-based estimators and show that a simple class of them is



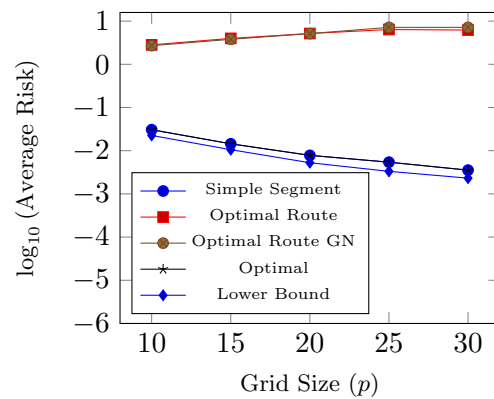
(a) $N = p^{1.0}$



(b) $N = p^{2.0}$

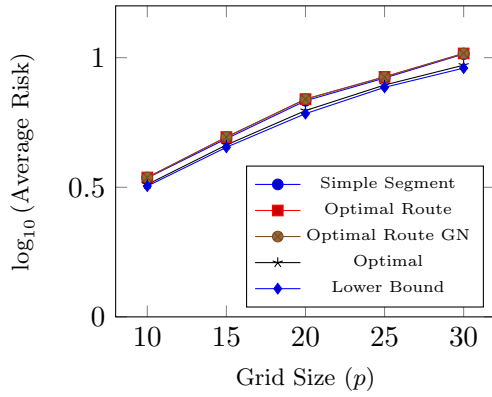


(c) $N = p^{3.0}$

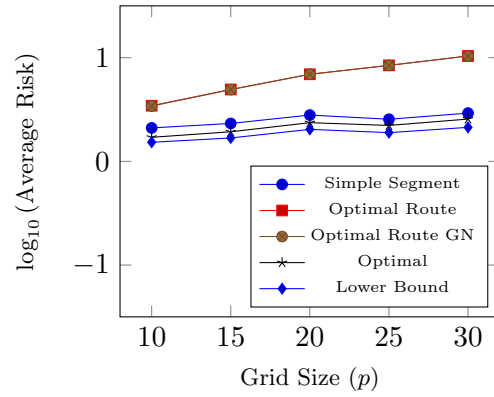


(d) $N = p^{4.0}$

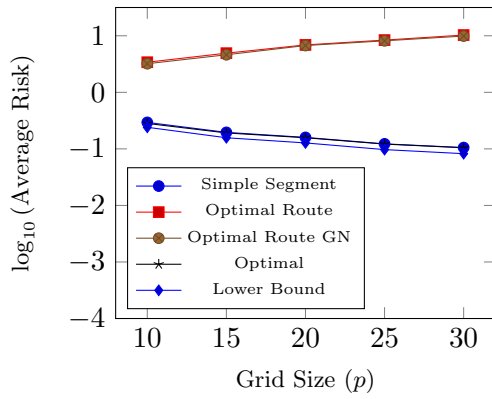
Figure 3: Average integrated risks of different methods ($\alpha = 0.3$).



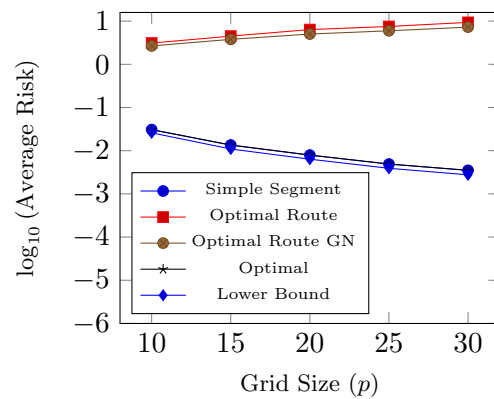
(a) $N = p^{1.0}$



(b) $N = p^{2.0}$

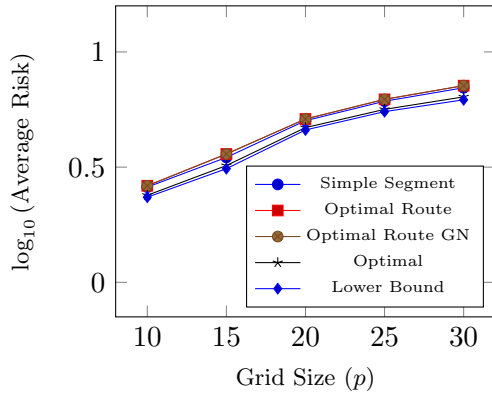


(c) $N = p^{3.0}$

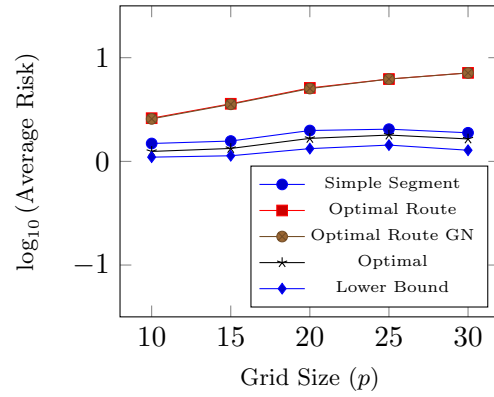


(d) $N = p^{4.0}$

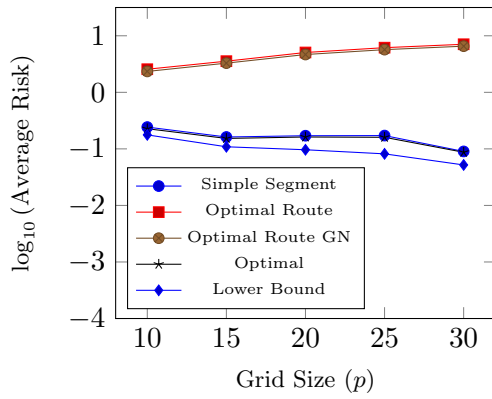
Figure 4: Average integrated risks of different methods ($\alpha = 1.0$).



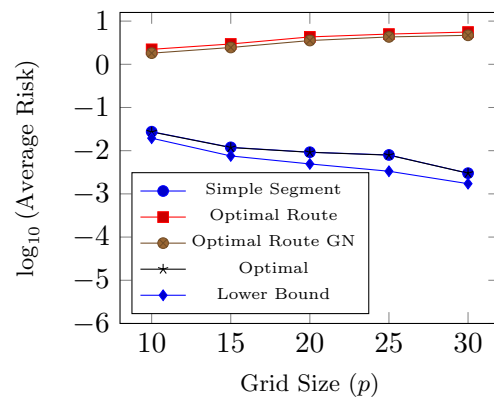
(a) $N = p^{1.0}$



(b) $N = p^{2.0}$



(c) $N = p^{3.0}$



(d) $N = p^{4.0}$

Figure 5: Average integrated risks of different methods ($\alpha = 3.0$).

asymptotically optimal up to a logarithmic factor with a variety of trip-generating processes on a grid network. At the core of our analysis is the following tradeoff. Segment-based estimators have the advantage of a larger sample size as there are more individual traversals on a segment level. However, the estimation can accumulate errors due to aggregating over road segments. On the other hand, route-based or generalized segment-based estimators can have the advantage of absorbing errors among segment travel times, but it is often at the cost of a smaller sample size. Our results expose that, under mild conditions, the sample size difference is often of first-order importance, leading to favorable consideration towards a segment-based approach.

It remains open whether similar insights hold under the setting of ETA prediction where one is only interested in predicting travel time from an origin to a destination without conditional on a route. Such settings occur in practice, for example, when one has little control over the route a driver will take [Hu et al., 2022]. Route-based methods which use data for all trip observations between the origin-destination pair can estimate travel time and the uncertain route distribution simultaneously, while segment-based methods require additional steps to estimate such route distribution. Extending our analyses in such settings can be meaningful follow-up work.

References

- James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013a.
- James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013b.
- Michael Bernard, Jeremy Keith Hackney, and Kay W. Axhausen. Correlation of link travel speeds. In *6th Swiss Transport Research Conference*, 2006.
- Min-Te Chao and WE Strawderman. Negative moments of positive random variables. *Journal of the American Statistical Association*, 67(338):429–431, 1972.
- DeepMind. Traffic prediction with advanced graph neural networks, 2020. URL <https://www.deepmind.com/blog/traffic-prediction-with-advanced-graph-neural-networks>.
- Daniel Delling, Andrew V Goldberg, Thomas Pajor, and Renato F Werneck. Customizable route planning in road networks. *Transportation Science*, 51(2):566–591, 2017.
- Austin Derrow-Pinion, Jennifer She, David Wong, Oliver Lange, Todd Hester, Luis Perez, Marc Nunkesser, Seongjae Lee, Xueying Guo, Brett Wiltshire, et al. ETA prediction with graph neural networks in Google Maps. *arXiv preprint arXiv:2108.11482*, 2021.
- DLMF. *NIST Digital Library of Mathematical Functions*. <http://dlmf.nist.gov/>, Release 1.0.28 of 2020-09-15, 2020. URL <http://dlmf.nist.gov/>.
- Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.
- Richard D Gill and Boris Y Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59–79, 1995.
- Feng Guo, Xin Gu, Zhaoxia Guo, Yucheng Dong, and Stein W Wallace. Understanding the marginal distributions and correlations of link travel speeds in road networks. *Scientific Reports*, 10(1): 1–8, 2020.

- Aude Hofleitner, Ryan Herring, Pieter Abbeel, and Alexandre Bayen. Learning the dynamics of arterial traffic from probe data using a dynamic Bayesian network. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1679–1693, 2012.
- Xinyu Hu, Tanmay Binaykiya, Eric Frank, and Olcay Cirit. DeepETA: An ETA post-processing system at scale. *arXiv preprint arXiv:2206.02127*, 2022.
- Erik Jenelius and Haris N Koutsopoulos. Travel time estimation for urban road networks using low frequency probe vehicle data. *Transportation Research Part B: Methodological*, 53:64–81, 2013.
- Ishan Jindal, Xuewen Chen, Matthew Nokleby, Jieping Ye, et al. A unified neural network approach for estimating travel time and distance for a taxi trip. *arXiv preprint arXiv:1710.04350*, 2017.
- Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.
- Yaguang Li, Kun Fu, Zheng Wang, Cyrus Shahabi, Jieping Ye, and Yan Liu. Multi-task representation learning for travel time estimation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1695–1704, 2018.
- José Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS*. Springer science & business media, 2006.
- Mohammed A Quddus, Washington Y Ochieng, and Robert B Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328, 2007.
- Piotr Rachtan, He Huang, and Song Gao. Spatiotemporal link speed correlations: Empirical study. *Transportation Research Record*, 2390(1):34–43, 2013.
- Mohsen Ramezani and Nikolas Geroliminis. On the estimation of arterial route travel time distribution with markov chains. *Transportation Research Part B: Methodological*, 46(10):1576–1590, 2012.
- Harry L van Trees. *Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory*. John Wiley & Sons, 2004.
- Hongjian Wang, Yu-Hsuan Kuo, Daniel Kifer, and Zhenhui Li. A simple baseline for travel time estimation using large-scale trip data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–4, 2016.
- Yilun Wang, Yu Zheng, and Yexiang Xue. Travel time estimation of a path using sparse trajectories. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 25–34, 2014.
- Zheng Wang, Kun Fu, and Jieping Ye. Learning to estimate the travel time. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 858–866, 2018a.
- Zheng Wang, Kun Fu, and Jieping Ye. Learning to estimate the travel time. In *KDD*, pages 858–866. ACM Special Interest Group on Knowledge Discovery and Data Mining, 2018b.

- Dawn Woodard, Galina Nogin, Paul Koch, David Racz, Moises Goldszmidt, and Eric Horvitz. Predicting travel time reliability using mobile phone GPS data. *Transportation Research Part C: Emerging Technologies*, 75:30–44, 2017.
- Chiwei Yan, Helin Zhu, Nikita Korolko, and Dawn Woodard. Dynamic pricing and matching in ride-hailing platforms. *Naval Research Logistics (NRL)*, 67(8):705–724, 2020.
- Haitao Yuan, Guoliang Li, Zhifeng Bao, and Ling Feng. Effective travel time estimation: When historical trajectories over road networks matter. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, pages 2135–2149, 2020.

Appendix

A Additional Technical Results

In this section, we give additional finite-sample and asymptotic results that are either used in or complement the main text. Their corresponding proofs can be found in Appendix B.

A.1 Finite Sample Results

Our first result in this subsection bounds the integrated risk of any estimator using a Bayesian information-theoretical bound. This lower bound is obtained by adapting the Bayesian Cramér-Rao bound [Gill and Levit, 1995] through the van Trees inequality [van Trees, 2004], a Bayesian analog of the information inequality. The proof uses Gaussian priors and posteriors to obtain explicit closed-form bounds.

Lemma 1 (INFORMATION-THEORETIC LOWER BOUND). In addition to Assumptions 1 and 3, assume that (\mathcal{E}, θ) are jointly Gaussian distributed. Given a set of historical routes $y_{[N]}$ and the predicting route y ,

$$R\left(\hat{\Theta}_y^* \mid y_{[N]}\right) \geq \frac{|y|^2}{\sum_{s,t \in y} N_{s \cup t} \psi_{s,t} + |y|/\tau^2}, \quad (11)$$

where $\psi_{s,t}$ is the $(s, t)^{\text{th}}$ element in the precision matrix of the segment travel times Ψ .

A.2 Asymptotic Results

We begin with a few remarks on the notation used in this section: for two functions $f(p)$ and $g(p) > 0$, we write $f(p) = \mathcal{O}(g(p))$ (or $f(p) = \Omega(g(p))$) if there exists a constant c_1 and a constant p_1 such that $f(p) \leq c_1 g(p)$ (or $f(p) \geq c_1 g(p)$) for all $p \geq p_1$; we write $f(p) = o(g(p))$ (or $f(p) = \omega(g(p))$) if $\lim_{p \rightarrow \infty} f(p)/g(p) = 0$ (or $\lim_{p \rightarrow \infty} f(p)/g(p) = +\infty$). In addition, we write $f(p) \gtrsim g(p)$ (or $f(p) \lesssim g(p)$) if there is a universal constant $c > 0$ such that $f(p) \geq c g(p)$ (or $f(p) \leq c g(p)$) for all $p \geq 1$. If $f(p) \lesssim g(p)$ and $f(p) \gtrsim g(p)$, we define $f(p) \simeq g(p)$.

Our first asymptotic result in this subsection generalizes Theorem 3 to non-grid road networks and general route distribution μ_p . The key quantities determining the integrated risks of these estimators are the rates at which training data accumulates on particular road segments or on particular routes. Under a road network with size p , we let $q_s := \mathbb{P}[s \in Y_p]$ denote the probability that a specific road segment s is traversed by a randomly generated route $Y_p \sim \mu_p$. Similarly, given any route y , we define $q_{\delta(y)} := \mathbb{P}[Y_p \in \delta(y)]$ as the probability that a randomly sampled route belongs to the neighborhood $\delta(y)$ of a given route y . With a slight abuse of notation, we further let $q_\delta := \mathbb{P}_{Y'_p \sim \mu_p, Y_p \sim \mu_p}[Y'_p \in \delta(Y_p)] = \sum_{y \in \mathcal{Y}_p} q_{\delta(y)} \mathbb{P}[Y_p = y]$. The quantity q_δ marginalizes over the distribution of the predicting route y and is the probability that a randomly sampled route Y'_p belongs to the neighborhood of another independently sampled route Y_p . Intuitively, q_δ measures the rate at which the neighborhood of any randomly sampled predicting route accumulates training data.

Our results will compare the asymptotic integrated risk of travel time estimators $R(\hat{\Theta}_{Y_p})$ as $p \rightarrow \infty$ (and $N \rightarrow \infty$). In our first asymptotic result below, we compare a large family of simple segment-based estimators $\hat{\Theta}_y^{(\text{seg})}$ to the optimal route-based estimators $\hat{\Theta}_y^{*(\text{route})}$. This family of

simple segment-based estimators only requires that $\phi_s(N_s)$ approaches 1 quickly enough. We give conditions under which this family of simple segment-based estimators is (much) more accurate when the size of the road network gets larger. This automatically implies that the optimal segment-based estimator also dominates the optimal route-based estimator, under the same set of conditions. We provide this result without restricting to non-negative covariances, as required in Theorem 2.

Theorem 5. *For any segment-based estimator $\hat{\Theta}_y^{(\text{seg})}$ with $\phi_s(N_s) = 1 - \mathcal{O}(1/\sqrt{N_s})$, $s \in \mathcal{S}_p$, and any optimal route-based estimator $\hat{\Theta}_y^{*(\text{route})}$ with neighborhood $\delta(\cdot)$ and route distribution μ_p such that $q_\delta = o(1/|S_p|)$,*

$$\lim_{p \rightarrow \infty} \frac{R\left(\hat{\Theta}_{Y_p}^{(\text{seg})}\right)}{R\left(\hat{\Theta}_{Y_p}^{*(\text{route})}\right)} = 0.$$

In addition, $R(\hat{\Theta}_{Y_p}^{(\text{seg})}) = \mathcal{O}(|S_p|/N)$ while $R(\hat{\Theta}_{Y_p}^{(\text{route})}) = \Omega(1/(Nq_\delta))$.*

Theorem 5 characterizes conditions under which a wide class of simple segment-based estimators dominates the optimal route-based estimator with a neighborhood such that the probability of any route within the neighborhood of a randomly sampled route being sampled scales as $o(1/|S_p|)$. We will give more explanation to this scaling under a grid network example in Section 3.1. It is interesting to note that Theorem 5 does not require any conditions on q_s . In the proof of Theorem 5, we lower bound the integrated risk of the optimal route-based estimator $R(\hat{\Theta}_{Y_p}^{*(\text{route})})$ by assuming that there is no additional bias introduced by including historical trips whose routes are not exactly the same as y into the neighborhood $\delta(y)$.

A.2.1 Asymptotic Results for the Grid Networks

We provide a few additional asymptotic results for the grid networks based on the route distribution described in Section 3.1. These are useful results that help to prove the main results Theorem 3 and Theorem 4. They concern various data accumulation rates on the grid networks. The first lemma below bounds the probability that a specific origin or destination is chosen on the grid.

Lemma 2. With $0 < \alpha \leq 1$,

$$p^{-2} \lesssim \mathbb{P}[X = (i, j)] \lesssim p^{-2\alpha}, \quad \forall (i, j) \in \mathcal{V}_p.$$

These bounds are tight. As we will show in the proof, at the four corners, $\mathbb{P}[X = (0, p)] = \mathbb{P}[X = (p, 0)] = \mathbb{P}[X = (0, 0)] = \mathbb{P}[X = (p, p)] \simeq p^{-2\alpha}$, while at the center of the grid we have $\mathbb{P}[X = ([p/2], [p/2])] \simeq p^{-2}$. With this lemma, we now give another result that characterizes the rate of q_δ , the probability that a randomly sampled route belongs to the neighborhood of another independently sampled route. In particular, we consider a neighborhood $\delta^{\text{od}}(\cdot)$ that includes all historical routes whose origins and destinations are close to those of the predicting route respectively. We define $x_1(y_p)$ and $x_2(y_p)$ as the origin and destination of route y_p . Construct $\delta^{\text{od}}(y_p) = \{y \in \mathcal{Y}_p : \|x_1(y), x_1(y_p)\|_1 \leq c, \|x_2(y), x_2(y_p)\|_1 \leq c\}$ for some fixed constant $c > 0$ that does *not* depend on p .

Proposition 4. Consider a route neighborhood $\delta^{\text{od}}(\cdot)$ that includes routes with similar origin and destination as those of the predicting route,

$$q_{\delta^{\text{od}}} = \mathbb{P}_{Y_p' \sim \mu_p, Y_p \sim \mu_p}[Y_p' \in \delta^{\text{od}}(Y_p)] = \sum_{y \in \mathcal{Y}_p} q_{\delta^{\text{od}}}(y) \cdot \mathbb{P}[Y_p = y] \simeq \begin{cases} p^{-4}, & 1/2 < \alpha \leq 1, \\ p^{-8\alpha}, & 0 < \alpha \leq 1/2. \end{cases}$$

Under route distribution μ_p specified in Figure 2 , we give bounds on the probability that a road segment s is traversed.

Proposition 5. Let $Y_p \sim \mu_p$. For any road segment $s \in \mathcal{S}_p$ in the grid,

$$p^{-1-\alpha} \lesssim q_s = \mathbb{P}[s \in Y_p] \lesssim p^{-\alpha}.$$

Figure 6 illustrates the result whose proof is provided in the supplement. For segments with horizontal orientation, the segments that accumulate the most amount of data are at the center of the upper and lower boundaries. On the other hand, the segments that accumulate the least amount of data lie at the center of the left and right boundaries. The data accumulation rates on segments with vertical orientation can be obtained by symmetry.

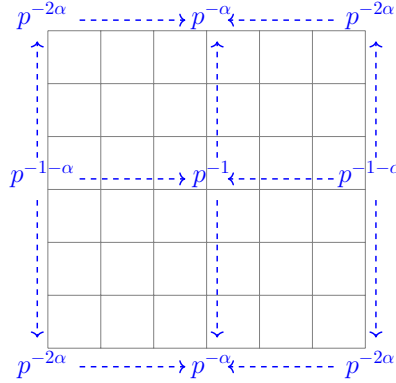


Figure 6: Data accumulation rates of traversals on segments with horizontal movements in a grid. The dashed arrows represent directions toward which the data accumulation rates over the segment increase.

B Proofs

Proof of Theorem 1. In the proof, we use a notation that is common in multivariate statistics. When we want to denote that a multivariate normal random vector Y has mean ν and covariance Σ , we will write

$$Y \sim \mathcal{N}(\nu, \Sigma).$$

Recall that $\mathcal{E} \in \mathbb{R}^{M \times 1}$ is the vector of error terms of all segment travel times. Specifically, \mathcal{E}_i is the error term of Z_i . With this notation, the entire data vector Z satisfies

$$Z \mid \theta \sim \mathcal{N}(U\theta, \Phi).$$

Let $b = \mu e$ and $B = \text{diag}(\tau^2 e)$.

$$\theta \sim \mathcal{N}(b, B).$$

First, notice that *marginal of θ* , we have

$$Z \sim \mathcal{N}(Ub, UBU^\top + \Phi).$$

This is because $\mathbb{E}[Z] = \mathbb{E}[\mathbb{E}[Z|\theta]] = U\mathbb{E}[\theta] = Ub$ and $\text{var}[Z] = \text{var}[\mathbb{E}[Z|\theta]] + \mathbb{E}[\text{var}[Z|\theta]] = UBU^\top + \Phi$. This means that jointly

$$\begin{pmatrix} Z \\ \theta \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} Ub \\ b \end{pmatrix}, \begin{pmatrix} UBU^\top + \Phi & A \\ A^\top & B \end{pmatrix} \right),$$

for some co-variance matrix A . This is because (\mathcal{E}, θ) are jointly Gaussian, $(U\theta + \mathcal{E}, \theta)^\top$, which is a linear transformation of $(\mathcal{E}, \theta)^\top$, must also be jointly Gaussian. Recall that, $Z = U\theta + \mathcal{E}$. Then, we have

$$A = \text{cov}(Z, \theta) = \text{cov}(U\theta + \mathcal{E}, \theta) = \text{cov}(U\theta, \theta) = UB,$$

where the second to the last equality holds by the independence of θ and \mathcal{E} .

We now know the entire covariance structure for the joint distribution of (Z, θ) . Then, the conditional distribution density $f(\theta | Z)$ should also be Gaussian. We have

$$\begin{aligned} & \log f(\theta | Z) \\ &= \log \frac{f(Z | \theta)f(\theta)}{f(Z)} \\ &= \log f(Z | \theta) + \log f(\theta) + c \\ &= -\frac{1}{2}(Z - U\theta)^\top \Phi^{-1}(Z - U\theta) - \frac{1}{2}(\theta - b)^\top B^{-1}(\theta - b) + c - \log \left(\sqrt{(2\pi)^M |\Phi|} \right) \end{aligned} \quad (12)$$

$$= -\frac{1}{2}(\theta - \mu_{\theta|Z})^\top \Sigma_{\theta|Z}^{-1}(\theta - \mu_{\theta|Z}) + c', \quad (13)$$

where c and c' are some constants, and in eq. (13) we utilize the fact that the conditional distribution density must be Gaussian, and $\mu_{\theta|Z}$ is the conditional expectation and $\Sigma_{\theta|Z}$ is the conditional covariance matrix.

Expanding eq. (12), the terms related to θ read

$$-\frac{1}{2} \left(\theta^\top (U^\top \Phi^{-1} U + B^{-1}) \theta - (Z^\top \Phi^{-1} U + b^\top B^{-1}) \theta - \theta^\top (U^\top \Phi^{-1} Z + B^{-1} b) \right).$$

Similarly, expanding eq. (13), the terms related to θ read

$$-\frac{1}{2} \left(\theta^\top \Sigma_{\theta|Z}^{-1} \theta - (\mu_{\theta|Z}^\top \Sigma_{\theta|Z}^{-1}) \theta - \theta^\top (\Sigma_{\theta|Z}^{-1} \mu_{\theta|Z}) \right).$$

To make them equal, we must have

$$\Sigma_{\theta|Z}^{-1} = U^\top \Phi^{-1} U + B^{-1}, \quad \mu_{\theta|Z} = \Sigma_{\theta|Z} (U^\top \Phi^{-1} Z + B^{-1} b).$$

For brevity of the notation, we let $Q := \Sigma_{\theta|Z}^{-1}$.

Because the estimator that minimizes the integrated risk based on squared error is the posterior mean [Berger, 2013b],

$$\hat{\theta}_y^* = \mathbb{E} \left[\sum_{s \in y} \theta_s \middle| Z \right] = e_y^\top Q^{-1} (U^\top \Phi^{-1} Z + B^{-1} b).$$

We now compute the integrated risk based on the squared error of the optimal estimator $\hat{\Theta}_y^*$. Recall that $E_y = e_y e_y^\top$. The integrated risk of $\hat{\Theta}_y^*$ conditional on θ is

$$\begin{aligned}
& \mathbb{E} \left[\left\| e_y^\top \theta - \hat{\Theta}_y^* \right\|^2 \middle| \theta \right] \\
&= \mathbb{E} \left[\left\| e_y^\top \theta - e_y^\top Q^{-1} (U^\top \Phi^{-1} Z + B^{-1} b) \right\|^2 \middle| \theta \right] \\
&= \mathbb{E} \left[\left\| e_y^\top (\theta - Q^{-1} B^{-1} b) - e_y^\top Q^{-1} U^\top \Phi^{-1} Z \right\|^2 \middle| \theta \right] \\
&= \mathbb{E} \left[(\theta - Q^{-1} B^{-1} b)^\top E_y (\theta - Q^{-1} B^{-1} b) - 2 (\theta - Q^{-1} B^{-1} b)^\top E_y Q^{-1} U^\top \Phi^{-1} Z \right. \\
&\quad \left. + Z^\top \Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top \Phi^{-1} Z \middle| \theta \right] \\
&= (\theta - Q^{-1} B^{-1} b)^\top E_y (\theta - Q^{-1} B^{-1} b) - 2 (\theta - Q^{-1} B^{-1} b)^\top E_y Q^{-1} U^\top \Phi^{-1} U \theta \\
&\quad + \text{tr} (\Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top \Phi^{-1} \Phi) + \theta^\top U^\top \Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top \Phi^{-1} U \theta \\
&= \theta^\top (E_y + U^\top \Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top \Phi^{-1} U - 2 E_y Q^{-1} U^\top \Phi^{-1} U) \theta + 2 b^\top B^{-1} Q^{-1} E_y Q^{-1} U^\top \Phi^{-1} U \theta \\
&\quad + \text{tr} (\Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top) + b^\top B^{-1} Q^{-1} E_y Q^{-1} B^{-1} b - 2 b^\top B^{-1} Q^{-1} E_y \theta.
\end{aligned}$$

Now we take an outer expectation over θ with respect to its prior to get the integrated risk,

$$\begin{aligned}
& \mathbb{E} \left[\mathbb{E} \left[\left\| e_y^\top \theta - \hat{\Theta}_y^* \right\|^2 \middle| \theta \right] \right] \\
&= \text{tr} (B (E_y + U^\top \Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top \Phi^{-1} U - 2 E_y Q^{-1} U^\top \Phi^{-1} U)) \\
&\quad + b^\top (E_y + U^\top \Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top \Phi^{-1} U - 2 E_y Q^{-1} U^\top \Phi^{-1} U) b \\
&\quad + 2 b^\top B^{-1} Q^{-1} E_y Q^{-1} U^\top \Phi^{-1} U b + \text{tr} (\Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top) \\
&\quad + b^\top B^{-1} Q^{-1} E_y Q^{-1} B^{-1} b - 2 b^\top B^{-1} Q^{-1} E_y b \\
&= \text{tr} (B (E_y + U^\top \Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top \Phi^{-1} U - 2 E_y Q^{-1} U^\top \Phi^{-1} U)) + \text{tr} (\Phi^{-1} U Q^{-1} E_y Q^{-1} U^\top).
\end{aligned}$$

This completes the proof. \square

Proof of Corollary 1. When $\sigma_{s,t} = 0$,

$$\begin{aligned}
Q &= U^\top \Phi^{-1} U + \text{diag} ((1/\tau^2) e) \\
&= \text{diag} \left([N_s/\sigma_s^2]_{s \in \mathcal{S}} \right) + \text{diag} ((1/\tau^2) e) \\
&= \text{diag} \left([N_s/\sigma_s^2 + 1/\tau^2]_{s \in \mathcal{S}} \right).
\end{aligned}$$

Moreover,

$$U^\top \Phi^{-1} z = \text{diag} \left(\left[\frac{\sum_{n: s \in y_n} T'_{n,s}}{\sigma_s^2} \right]_{s \in \mathcal{S}} \right).$$

This gives,

$$\begin{aligned}
\hat{\Theta}_y^* &= e_y^\top Q^{-1} (U^\top \Phi^{-1} z + (\mu/\tau^2) e) \\
&= e_y^\top \text{diag} \left(\left[(N_s/\sigma_s^2 + 1/\tau^2)^{-1} \right]_{s \in \mathcal{S}} \right) \text{diag} \left(\left[\sum_{n:s \in y_n} T'_{n,s}/\sigma_s^2 + \mu/\tau^2 \right]_{s \in \mathcal{S}} \right) \\
&= e_y^\top \text{diag} \left(\left[\frac{\sum_{n:s \in y_n} T'_{n,s}/\sigma_s^2 + \mu/\tau^2}{N_s/\sigma_s^2 + 1/\tau^2} \right]_{s \in \mathcal{S}} \right) \\
&= \sum_{s \in y} \left(\frac{\sigma_s^2}{N_s \tau^2 + \sigma_s^2} \cdot \mu + \frac{N_s \tau^2}{N_s \tau^2 + \sigma_s^2} \cdot \frac{\sum_{n:s \in y_n} T'_{n,s}}{N_s} \right).
\end{aligned}$$

This completes the proof. \square

Proof of Proposition 1. For the generalized segment-based estimator $\hat{\Theta}_y^{(\text{g-seg})}$,

$$\begin{aligned}
& \mathbb{E} \left[\left(\hat{\Theta}_y^{(\text{g-seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right] \\
&= \mathbb{E} \left[\mathbb{E} \left[\left(\hat{\Theta}_y^{(\text{g-seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| \{\theta_s\}_{s \in y}, y_{[N]} \right] \middle| y_{[N]} \right] \\
&= \mathbb{E} \left[\text{var} \left(\hat{\Theta}_y^{(\text{g-seg})} \middle| \{\theta_s\}_{s \in y}, y_{[N]} \right) + \text{Bias}^2 \left(\hat{\Theta}_y^{(\text{g-seg})} \middle| \{\theta_s\}_{s \in y}, y_{[N]} \right) \middle| y_{[N]} \right] \\
&= \mathbb{E} \left[\text{var} \left(\sum_{S \in \mathcal{S}_y} \phi_S(N_S) \frac{\sum_{n:S \subset y_n} \sum_{s \in S} T'_{n,s}}{N_S} \middle| \{\theta_s\}_{s \in y}, y_{[N]} \right) + \left(\sum_{S \in \mathcal{S}_y} (1 - \phi_S(N_S)) \left(|S|\mu - \sum_{s \in S} \theta_s \right) \right)^2 \middle| y_{[N]} \right] \\
&= \mathbb{E} \left[\sum_{S,T \in \mathcal{S}_y} \frac{N_{S \cup T}}{N_S N_T} \phi_S(N_S) \phi_T(N_T) \left(\sum_{s \in S, t \in T} \sigma_{s,t} \right) + \left(\sum_{S \in \mathcal{S}_y} (1 - \phi_S(N_S)) \left(|S|\mu - \sum_{s \in S} \theta_s \right) \right)^2 \middle| y_{[N]} \right] \\
&= \sum_{S,T \in \mathcal{S}_y} \frac{N_{S \cup T}}{N_S N_T} \phi_S(N_S) \phi_T(N_T) \left(\sum_{s \in S, t \in T} \sigma_{s,t} \right) + \mathbb{E} \left[\left(\sum_{S \in \mathcal{S}_y} (1 - \phi_S(N_S)) \left(|S|\mu - \sum_{s \in S} \theta_s \right) \right)^2 \middle| y_{[N]} \right] \\
&= \sum_{S,T \in \mathcal{S}_y} \frac{N_{S \cup T}}{N_S N_T} \phi_S(N_S) \phi_T(N_T) \left(\sum_{s \in S, t \in T} \sigma_{s,t} \right) + \text{var} \left(\sum_{S \in \mathcal{S}_y} (1 - \phi_S(N_S)) \left(|S|\mu - \sum_{s \in S} \theta_s \right) \middle| y_{[N]} \right) \\
&= \sum_{S,T \in \mathcal{S}_y} \frac{N_{S \cup T}}{N_S N_T} \phi_S(N_S) \phi_T(N_T) \left(\sum_{s \in S, t \in T} \sigma_{s,t} \right) + \sum_{S \in \mathcal{S}_y} (1 - \phi_S(N_S))^2 |S| \tau^2.
\end{aligned}$$

Similarly, for the route-based estimator $\hat{\Theta}_y^{(\text{route})}$, to simplify the notation, we define $\Delta_{\delta(y)} = (\sum_{n:y_n \in \delta(y)} \sum_{s \in y_n} \theta_s - M_{\delta(y)} \sum_{s \in y} \theta_s) / M_{\delta(y)}$. We have the following risk calculation.

$$\mathbb{E} \left[\left(\hat{\Theta}_y^{(\text{route})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right]$$

$$\begin{aligned}
&= \mathbb{E} \left[\text{var} \left(\hat{\Theta}_y^{(\text{route})} \mid \{\theta_s\}_{s \in y}, y_{[N]} \right) + \text{Bias}^2 \left(\hat{\Theta}_y^{(\text{route})} \mid \{\{\theta_s\}_{s \in y}, y_{[N]}\} \right) \mid y_{[N]} \right] \\
&= \mathbb{E} \left[\text{var} \left(\phi_{\delta(y)}(M_{\delta(y)}) \frac{\sum_{n: y_n \in \delta(y)} \sum_{s \in y} T'_{n,s}}{M_{\delta(y)}} \mid \{\theta_s\}_{s \in y}, y_{[N]} \right) \right. \\
&\quad \left. + \left(\phi_{\delta(y)}(M_{\delta(y)}) \Delta_{\delta(y)} + \sum_{s \in y} (1 - \phi_{\delta(y)}(M_{\delta(y)})) (\mu - \theta_s) \right)^2 \mid y_{[N]} \right] \\
&= \mathbb{E} \left[\left(\frac{\phi_{\delta(y)}(M_{\delta(y)})}{M_{\delta(y)}} \right)^2 \left(\sum_{n: y_n \in \delta(y)} \sum_{s, t \in y_n} \sigma_{s,t} \right) \right. \\
&\quad \left. + \left(\phi_{\delta(y)}(M_{\delta(y)}) \Delta_{\delta(y)} + \sum_{s \in y} (1 - \phi_{\delta(y)}(M_{\delta(y)})) (\mu - \theta_s) \right)^2 \mid y_{[N]} \right] \\
&= \left(\frac{\phi_{\delta(y)}(M_{\delta(y)})}{M_{\delta(y)}} \right)^2 \left(\sum_{n: y_n \in \delta(y)} \sum_{s, t \in y_n} \sigma_{s,t} \right) + (\phi_y(M_{\delta(y)}) \mathbb{E}[\Delta_{\delta(y)}])^2 \\
&\quad + \text{var} \left(\phi_{\delta(y)}(M_{\delta(y)}) \Delta_{\delta(y)} + \sum_{s \in y} (1 - \phi_{\delta(y)}(M_{\delta(y)})) (\mu - \theta_s) \mid y_{[N]} \right).
\end{aligned}$$

We further have,

$$\mathbb{E}[\Delta_{\delta(y)}] = \frac{\left(\sum_{n: y_n \in \delta(y)} |y_n| \right) - M_{\delta(y)} |y|}{M_{\delta(y)}} \mu = (\bar{y}_{\delta(y)} - |y|) \mu,$$

and,

$$\begin{aligned}
&\text{var} \left(\sum_{s \in y} (1 - \phi_{\delta(y)}(M_{\delta(y)})) (\mu - \theta_s) + \phi_{\delta(y)}(M_{\delta(y)}) \Delta_{\delta(y)} \mid y_{[N]} \right) \\
&= \text{var} \left(|y| (1 - \phi_y(M_{\delta(y)})) \mu - (1 - \phi_{\delta(y)}(M_{\delta(y)})) \left(\sum_{s \in y} \theta_s \right) \right. \\
&\quad \left. - \phi_{\delta(y)}(M_{\delta(y)}) \left(\sum_{s \in y} \theta_s \right) + \phi_{\delta(y)}(M_{\delta(y)}) \frac{\sum_{n: y_n \in \delta(y)} \sum_{s \in y_n} \theta_s}{M_{\delta(y)}} \right) \\
&= \text{var} \left(- \left(\sum_{s \in y} \theta_s \right) + \phi_{\delta(y)}(M_{\delta(y)}) \frac{\sum_{n: y_n \in \delta(y)} \sum_{s \in y_n} \theta_s}{M_{\delta(y)}} \right) \\
&= \text{var} \left(- \left(\sum_{s \in y} \theta_s \right) + \phi_{\delta(y)}(M_{\delta(y)}) \left(\sum_{s \in y} \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \theta_s + \sum_{s \in \mathcal{S}_{\delta(y)} \setminus y} \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \theta_s \right) \right) \\
&= \sum_{s \in y} \left(1 - \phi_{\delta(y)}(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2 + \sum_{s \in \mathcal{S}_{\delta(y)} \setminus y} \left(\phi_{\delta(y)}(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2.
\end{aligned}$$

Putting this all together leads to,

$$R \left(\hat{\Theta}_y^{(\text{route})} \mid y_{[N]} \right) = \left(\frac{\phi_{\delta(y)}(M_{\delta(y)})}{M_{\delta(y)}} \right)^2 \left(\sum_{n: y_n \in \delta(y)} \sum_{s, t \in y_n} \sigma_{s,t} \right) + (\phi_{\delta(y)}(M_{\delta(y)}) (\bar{y}_{\delta(y)} - |y|) \mu)^2$$

$$\begin{aligned}
& + \sum_{s \in \mathcal{S}_{\delta(y)} \setminus y} \left(\phi_{\delta(y)}(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2 + \sum_{s \in y} \left(1 - \phi_{\delta(y)}(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2 \\
& = \left(\frac{\phi_{\delta(y)}(M_{\delta(y)})}{M_{\delta(y)}} \right)^2 \left(\sum_{s,t \in \mathcal{S}_{\delta(y)}} N_{s \cup t}^{\delta(y)} \sigma_{s,t} \right) + (\phi_{\delta(y)}(M_{\delta(y)}) (\bar{y}_{\delta(y)} - |y|) \mu)^2 \\
& + \sum_{s \in \mathcal{S}_{\delta(y)} \setminus y} \left(\phi_{\delta(y)}(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2 + \sum_{s \in y} \left(1 - \phi_{\delta(y)}(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2.
\end{aligned}$$

This completes the proof. \square

Proof of Proposition 2. For any route y ,

$$\phi_{\delta(y)}^*(M_{\delta(y)}) = \arg \min_{\phi_y(\cdot)} \mathbb{E} \left[\left(\hat{\Theta}_y^{(\text{route})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right]$$

is well defined by checking the first-order condition of (4) as $\mathbb{E}[(\hat{\Theta}_y^{(\text{route})} - \sum_{s \in y} \theta_s)^2 | y_{[N]}]$ is strictly convex in $\phi_{\delta(y)}(M_{\delta(y)})$,

$$\begin{aligned}
& \frac{2\phi_{\delta(y)}^*(M_{\delta(y)})}{(M_{\delta(y)})^2} \left(\sum_{y_n \in \delta(y)} \sum_{s,t \in y_n} \sigma_{s,t} \right) + 2(\mu(\bar{y}_{\delta(y)} - |y|))^2 \phi_{\delta(y)}^*(M_{\delta(y)}) + \sum_{s \in \mathcal{S}_{\delta(y)} \setminus y} 2\tau^2 \left(\frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \phi_{\delta(y)}^*(M_{\delta(y)}) \\
& = \sum_{s \in y} 2\tau^2 \left(1 - \phi_{\delta(y)}^*(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right) \frac{N_s^{\delta(y)}}{M_{\delta(y)}}.
\end{aligned}$$

This gives the optimal route-based estimator $\hat{\Theta}_y^{*(\text{route})}$,

$$\hat{\Theta}_y^{*(\text{route})} := (1 - \phi_{\delta(y)}^*(M_{\delta(y)})) |y| \theta + \phi_{\delta(y)}^*(M_{\delta(y)}) \frac{\sum_{n: y_n = y} \sum_{s \in y} T'_{n,s}}{M_{\delta(y)}},$$

$$\phi_{\delta(y)}^*(M_{\delta(y)}) = \left(\sum_{s \in y} N_s^{\delta(y)} \right) \tau^2 / \left(\sum_{s \in \mathcal{S}_{\delta(y)}} \frac{(N_s^{\delta(y)})^2}{M_{\delta(y)}} \tau^2 + \frac{\sum_{n: y_n \in \delta(y)} \sum_{s,t \in y_n} \sigma_{s,t}}{M_{\delta(y)}} + M_{\delta(y)} \mu^2 (\bar{y}_{\delta(y)} - |y|)^2 \right).$$

It can be checked that the Hessian of the integrated risk $\mathbb{E}[(\hat{\Theta}_y^{(\text{g-seg})} - \sum_{s \in y} \theta_s)^2 | y_{[N]}]$ is symmetric and positive definite (PD). Suppose $|\mathcal{S}_y| = m$. Let S_i , $i \in \{1, \dots, m\}$ be the i^{th} super-segment in \mathcal{S}_y .

$$\begin{aligned}
& \text{Hess} \left(\mathbb{E} \left[\left(\hat{\Theta}_y^{(\text{g-seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right] \right) \\
& = 2 \cdot \begin{bmatrix} \frac{1}{N_{S_1}} \left(\sum_{s \in S_1, t \in S_1} \sigma_{s,t} \right) + |S_1| \tau^2 & \dots & \frac{N_{S_i \cup S_j}}{N_{S_i} N_{S_j}} \left(\sum_{s \in S_i, t \in S_j} \sigma_{s,t} \right) & \dots \\ \vdots & \ddots & \vdots & \dots \\ \frac{N_{S_i \cup S_j}}{N_{S_i} N_{S_j}} \left(\sum_{s \in S_i, t \in S_j} \sigma_{s,t} \right) & \dots & \dots & \dots \\ \vdots & \dots & \dots & \frac{1}{N_{S_m}} \left(\sum_{s \in S_m, t \in S_m} \sigma_{s,t} \right) + |S_m| \tau^2 \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
&= 2 \cdot \begin{bmatrix} \frac{1}{N_{S_1}} \left(\sum_{s \in S_1, t \in S_1} \sigma_{s,t} \right) & \cdots & \frac{N_{S_i \cup S_j}}{N_{S_i} N_{S_j}} \left(\sum_{s \in S_i, t \in S_j} \sigma_{s,t} \right) & \cdots \\ \vdots & \ddots & \vdots & \cdots \\ \frac{N_{S_i \cup S_j}}{N_{S_i} N_{S_j}} \left(\sum_{s \in S_i, t \in S_j} \sigma_{s,t} \right) & \cdots & \ddots & \cdots \\ \vdots & \cdots & \cdots & \frac{1}{N_{S_m}} \left(\sum_{s \in S_m, t \in S_m} \sigma_{s,t} \right) \end{bmatrix} \\
&+ 2\tau^2 \cdot \begin{bmatrix} |S_1| & & & \\ & \ddots & & \\ & & & \\ & & & |S_m| \end{bmatrix}.
\end{aligned}$$

The first matrix in the last equality is positive semidefinite (PSD). To see this, note that $1/N_{S_i} \geq N_{S_i \cup S_j} / (N_{S_i} N_{S_j})$, $\forall i, j \in \{1, \dots, m\}$. Such scaling of the entries in a PSD matrix (the covariance matrix of super-segment travel times is PSD) results in a PSD matrix. Moreover, the second matrix in the last equality is positive definite (PD) because it is a diagonal matrix with strictly positive entries. As the sum of PSD and PD matrices is PD, this verifies the strict convexity of the integrated risk. The first-order conditions, presented in the statement, must admit a unique solution. This completes the proof. \square

Proof of Theorem 2. The proof is by construction. Consider a segment-based estimator $\hat{\Theta}'_y^{(\text{seg})}$ with $\phi_s(N_s) = \phi_{\delta(y)}^*(M_{\delta(y)}) \cdot (N_s^{\delta(y)} / M_{\delta(y)})$ where $\phi_{\delta(y)}^*(\cdot)$ has a closed-form as indicated in Proposition 2. For any set of historical routes $y_{[N]}$, the integrated risk of this estimator is,

$$\begin{aligned}
&\mathbb{E} \left[\left(\hat{\Theta}'_y^{(\text{seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right] \\
&= \sum_{s,t \in y} \frac{N_{s \cup t}}{N_s N_t} \frac{N_s^{\delta(y)} N_t^{\delta(y)}}{M_{\delta(y)}^2} \phi_{\delta(y)}^*(M_{\delta(y)})^2 \sigma_{s,t} + \sum_{s \in y} \left(1 - \phi_{\delta(y)}^*(M_{\delta(y)}) \cdot \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2 \\
&\leq \frac{\phi_{\delta(y)}^*(M_{\delta(y)})^2}{M_{\delta(y)}^2} \left(\sum_{s,t \in y} N_{s \cup t}^{\delta(y)} \sigma_{s,t} \right) + \sum_{s \in y} \left(1 - \phi_{\delta(y)}^*(M_{\delta(y)}) \cdot \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2 \\
&\leq \frac{\phi_{\delta(y)}^*(M_{\delta(y)})^2}{M_{\delta(y)}^2} \left(\sum_{s,t \in \mathcal{S}_{\delta(y)}} N_{s \cup t}^{\delta(y)} \sigma_{s,t} \right) + \sum_{s \in y} \left(1 - \phi_{\delta(y)}^*(M_{\delta(y)}) \cdot \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2 \\
&\leq \left(\frac{\phi_{\delta(y)}^*(M_{\delta(y)})}{M_{\delta(y)}} \right)^2 \left(\sum_{s,t \in \mathcal{S}_{\delta(y)}} N_{s \cup t}^{\delta(y)} \sigma_{s,t} \right) + \left(\phi_{\delta(y)}^*(M_{\delta(y)}) (\bar{y}_{\delta(y)} - |y|) \mu \right)^2 \\
&\quad + \sum_{s \in \mathcal{S}_{\delta(y)} \setminus y} \left(\phi_{\delta(y)}^*(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2 + \sum_{s \in y} \left(1 - \phi_{\delta(y)}^*(M_{\delta(y)}) \frac{N_s^{\delta(y)}}{M_{\delta(y)}} \right)^2 \tau^2 \\
&= \mathbb{E} \left[\left(\hat{\Theta}^{*(\text{route})}_y - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right].
\end{aligned}$$

The first inequality uses the assumption that $N_{s \cup t} N_s^{\delta(y)} N_t^{\delta(y)} \leq N_{s \cup t}^{\delta(y)} N_s N_t$. The second inequality uses $\sigma_{s,t} \geq 0$, $\forall s, t \in \mathcal{S}$ and $\mathcal{S}_{\delta(y)} \supset y$. The third inequality holds because the additional two terms

(the second and third terms) are both non-negative. The proof is then completed by

$$\mathbb{E} \left[\left(\hat{\Theta}_y^{*(\text{seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right] \leq \mathbb{E} \left[\left(\hat{\Theta}'_y^{(\text{seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right] \leq \mathbb{E} \left[\left(\hat{\Theta}_y^{*(\text{route})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right].$$

□

Proof of Proposition 3. We first show that $R(\hat{\Theta}_y^{*(\text{seg})} | y_{[N]}) \leq R(\hat{\Theta}_y^{*(\text{g-seg})} | y_{[N]})$. Consider a segment-based estimator $\hat{\Theta}_y^{(\text{seg})}$ with $\phi_s(N_s) = \phi_y^*(N_y)$, $\forall s \in y$. Note that here $\phi_y^*(\cdot)$ has a closed form $\phi_y^*(N_y) = N_y |y| \tau^2 / (N_y |y| \tau^2 + \sum_{s,t \in y} \sigma_{s,t})$ as $\mathcal{S}_y = \{\{y\}\}$. For any set of historical routes $y_{[N]}$, the integrated risk of this estimator is,

$$\begin{aligned} & \mathbb{E} \left[\left(\hat{\Theta}_y^{(\text{seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right] \\ &= \sum_{s,t \in y} \frac{N_{s \cup t}}{N_s N_t} \phi_y^*(N_y)^2 \sigma_{s,t} + \sum_{s \in y} (1 - \phi_y^*(N_y))^2 \tau^2 \\ &\leq \sum_{s,t \in y} \frac{1}{N_y} \phi_y^*(N_y)^2 \sigma_{s,t} + |y| (1 - \phi_y^*(N_y))^2 \tau^2 \\ &= \mathbb{E} \left[\left(\hat{\Theta}_y^{*(\text{g-seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right]. \end{aligned} \tag{14}$$

Inequality (14) holds because $N_y \leq N_s$ and $N_{s \cup t} \leq N_t$ and the assumption that $\sigma_{s,t} \geq 0$. These imply $\sigma_{s,t} N_{s \cup t} / (N_s N_t) \leq \sigma_{s,t} / N_s \leq \sigma_{s,t} / N_y$. The proof of the inequality is then completed by

$$\mathbb{E} \left[\left(\hat{\Theta}_y^{*(\text{seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right] \leq \mathbb{E} \left[\left(\hat{\Theta}'_y^{(\text{seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right] \leq \mathbb{E} \left[\left(\hat{\Theta}_y^{*(\text{g-seg})} - \sum_{s \in y} \theta_s \right)^2 \middle| y_{[N]} \right].$$

We then prove that $R(\hat{\Theta}_y^{*(\text{g-seg})} | y_{[N]}) \leq R(\hat{\Theta}_y^{*(\text{route})} | y_{[N]})$. When $\mathcal{S}_y = \{\{y\}\}$ and $\delta(y) = \{y_n : y_n = y\}$, $\phi_y^*(\cdot)$ and $\phi_{\delta(y)}^*(\cdot)$ have the same form,

$$\phi_y^*(N_y) = \frac{N_y |y| \tau^2}{N_y |y| \tau^2 + \sum_{s,t \in y} \sigma_{s,t}}, \quad \phi_{\delta(y)}^*(M_{\delta(y)}) = \frac{M_{\delta(y)} |y| \tau^2}{M_{\delta(y)} |y| \tau^2 + \sum_{s,t \in y} \sigma_{s,t}}.$$

The optimal integrated risks also share the same form,

$$\begin{aligned} R(\hat{\Theta}_y^{*(\text{g-seg})} | y_{[N]}) &= \frac{1}{N_y} (\phi_y^*(N_y))^2 \left(\sum_{s,t \in y} \sigma_{s,t} \right) + (1 - \phi_y^*(N_y))^2 |y| \tau^2, \\ R(\hat{\Theta}_y^{*(\text{route})} | y_{[N]}) &= \frac{1}{M_{\delta(y)}} (\phi_{\delta(y)}^*(M_{\delta(y)}))^2 \left(\sum_{s,t \in y} \sigma_{s,t} \right) + (1 - \phi_{\delta(y)}^*(M_{\delta(y)}))^2 |y| \tau^2. \end{aligned}$$

Because $N_y \geq M_{\delta(y)}$ and the optimal integrated risk decreases with sample size, we thus have

$$R(\hat{\Theta}_y^{*(\text{g-seg})} | y_{[N]}) \leq R(\hat{\Theta}_y^{*(\text{route})} | y_{[N]}).$$

This completes the proof. □

Proof of Theorem 5. Under a given road network size p , let indicator variable I_s denote whether a road segment s is traversed from a randomly sampled route $Y_p \sim \mu_p$. For the segment-based estimator, given a predicting route that covers road segments indicated by $I = \{I_s\}_{s \in \mathcal{S}_p}$,

$$\begin{aligned}
R\left(\hat{\Theta}_{Y_p}^{(\text{seg})} \mid I\right) &= \sum_{s,t \in \mathcal{S}_p} \mathbb{E} \left[\frac{N_{s \cup t}}{N_s N_t} \phi_s(N_s) \phi_t(N_t) \right] I_s I_t \sigma_{s,t} + \sum_{s \in \mathcal{S}_p} \mathbb{E} [(1 - \phi_s(N_s))^2] I_s \tau^2 \\
&\leq \sum_{s,t \in \mathcal{S}_p} \mathbb{E} \left[\frac{N_{s \cup t}}{N_s N_t} \mathbf{1}\{N_s, N_t > 0\} \right] I_s I_t \max\{\sigma_{s,t}, 0\} + \sum_{s \in \mathcal{S}_p} \mathbb{E} [(1 - \phi_s(N_s))^2] I_s \tau^2 \\
&\leq \sum_{s,t \in \mathcal{S}_p} \mathbb{E} \left[\frac{N_{s \cup t}}{N_s N_t} \mathbf{1}\{N_s, N_t > 0\} \right] I_s I_t |\sigma_{s,t}| + \sum_{s \in \mathcal{S}_p} \mathbb{E} [(1 - \phi_s(N_s))^2] I_s \tau^2. \tag{15}
\end{aligned}$$

Note that $\mathbb{E}[N_s] = Nq_s$. By Chernoff bound, for any $\beta > 0$, $\mathbb{P}(N_s \leq (1 - \beta)\mathbb{E}[N_s]) = \mathbb{P}(N_s \leq (1 - \beta)Nq_s) \leq e^{-\beta^2 Nq_s/2}$. This yields,

$$\begin{aligned}
&\mathbb{E} [(1 - \phi_s(N_s))^2] \\
&= \mathbb{E} [(1 - \phi_s(N_s))^2 \mid N_s \leq (1 - \beta)Nq_s] \mathbb{P}(N_s \leq (1 - \beta)Nq_s) \\
&\quad + \mathbb{E} [(1 - \phi_s(N_s))^2 \mid N_s > (1 - \beta)Nq_s] \mathbb{P}(N_s > (1 - \beta)Nq_s) \\
&\leq \mathbb{P}(N_s \leq (1 - \beta)Nq_s) + \mathbb{E} [(1 - \phi_s(N_s))^2 \mid N_s > (1 - \beta)Nq_s] \\
&= \mathbb{P}(N_s \leq (1 - \beta)Nq_s) + \mathbb{E} [\mathcal{O}(1/N_s) \mid N_s > (1 - \beta)Nq_s] \\
&\leq \mathbb{P}(N_s \leq (1 - \beta)Nq_s) + \mathcal{O}(1/((1 - \beta)Nq_s)) \\
&= \mathcal{O}(1/(Nq_s)). \tag{16}
\end{aligned}$$

The second equality holds because $(1 - \phi_s(N_s)) = \mathcal{O}(1/\sqrt{N_s})$ and the last equality holds because $\mathbb{P}(N_s \leq (1 - \beta)Nq_s) \leq e^{-\beta^2 Nq_s/2} = \mathcal{O}(1/(Nq_s))$. Using this observation,

$$\begin{aligned}
(15) &= \sum_{s,t \in \mathcal{S}_p} \mathbb{E} \left[\frac{N_{s \cup t}}{N_s N_t} \mathbf{1}\{N_s, N_t > 0\} \right] I_s I_t |\sigma_{s,t}| + \sum_{s \in \mathcal{S}_p} \mathcal{O}(1/(Nq_s)) I_s \tau^2 \\
&\leq \sum_{s,t \in \mathcal{S}_p} \mathbb{E} \left[\frac{1}{N_{s \cup t}} \mathbf{1}\{N_{s \cup t} > 0\} \right] I_s I_t |\sigma_{s,t}| + \sum_{s \in \mathcal{S}_p} \mathcal{O}(1/(Nq_s)) I_s \tau^2 \\
&\stackrel{(\text{Lemma 3})}{\leq} \sum_{s,t \in \mathcal{S}_p} \frac{2}{\mathbb{E}[N_{s \cup t}]} I_s I_t |\sigma_{s,t}| + \sum_{s \in \mathcal{S}_p} \mathcal{O}(1/(Nq_s)) I_s \tau^2.
\end{aligned}$$

The second inequality holds as $N_{s \cup t} \leq N_s$ and $N_{s \cup t} \leq N_t$ for any $s, t \in \mathcal{S}_p$. The third equality holds by Lemma 3 (Appendix C) which implies that $\mathbb{E} [1/N_s \mathbf{1}\{N_s > 0\}] < 2/(q_s N)$.

Taking the expectation over $I = \{I_s\}_{s \in \mathcal{S}_p}$,

$$R\left(\hat{\Theta}_{Y_p}^{(\text{seg})}\right) \leq \sum_{s,t \in \mathcal{S}_p} \frac{2}{\mathbb{E}[N_{s \cup t}]} \mathbb{E} [I_s I_t] |\sigma_{s,t}| + \sum_{s \in \mathcal{S}_p} \mathcal{O}(1/(Nq_s)) \mathbb{E} [I_s] \tau^2.$$

Note that $\mathbb{E}[I_s I_t] = \mathbb{P}(I_s = 1, I_t = 1) = \mathbb{E}[N_{s \cup t}]/N$ and $\mathbb{E}[I_s] = \mathbb{E}[N_s]/N$. This gives,

$$R\left(\hat{\Theta}_{Y_p}^{(\text{seg})}\right) \leq \sum_{s,t \in \mathcal{S}_p} \frac{2}{\mathbb{E}[N_{s \cup t}]} \mathbb{E} [I_s I_t] |\sigma_{s,t}| + \sum_{s \in \mathcal{S}_p} \mathcal{O}(1/(Nq_s)) \mathbb{E} [I_s] \tau^2$$

$$\begin{aligned}
&= \sum_{s,t \in \mathcal{S}_p} \frac{2}{N} |\sigma_{s,t}| + \sum_{s \in \mathcal{S}_p} \mathcal{O}(1/N) \tau^2 \\
&= \mathcal{O}(|\mathcal{S}_p|/N).
\end{aligned} \tag{17}$$

The last equality uses the second part of Assumption 4. We now focus on the integrated risk of the optimal route-based estimator $\hat{\Theta}_{Y_p}^{*(\text{route})}$, conditional on the predicting route Y_p ,

$$\begin{aligned}
R\left(\hat{\Theta}_{Y_p}^{*(\text{route})} \mid Y_p\right) &= \mathbb{E} \left[\left(\frac{\phi_{Y_p}^*(M_{\delta(Y_p)})}{M_{\delta(Y_p)}} \right)^2 \sum_{n: y_n \in \delta(Y_p)} \sum_{s,t \in y_n} \sigma_{s,t} \right] + \mathbb{E} \left[\left(\phi_{Y_p}^*(M_{\delta(Y_p)}) (\bar{y}_{\delta(Y_p)} - |Y_p| \mu) \right)^2 \right] \\
&\quad + \sum_{s \in \delta(Y_p) \setminus Y_p} \mathbb{E} \left[\left(\frac{N_s^{\delta(Y_p)}}{M_{\delta(Y_p)}} \right)^2 \tau^2 \right] + \sum_{s \in Y_p} \mathbb{E} \left[\left(1 - \phi_{Y_p}^*(M_{\delta(Y_p)}) \frac{N_s^{\delta(Y_p)}}{M_{\delta(Y_p)}} \right)^2 \tau^2 \right] \\
&\geq \mathbb{E} \left[\left(\frac{\phi_{Y_p}^*(M_{\delta(Y_p)})}{M_{\delta(Y_p)}} \right)^2 M_{\delta(Y_p)} \sigma_{\min}^2 \right] + \sum_{s \in Y_p} \mathbb{E} \left[\left(1 - \phi_{Y_p}^*(M_{\delta(Y_p)}) \right)^2 \tau^2 \right] \\
&= \mathbb{E} \left[\frac{\phi_{Y_p}^*(M_{\delta(Y_p)})^2}{M_{\delta(Y_p)}} \right] \sigma_{\min}^2 + \sum_{s \in Y_p} \mathbb{E} \left[\left(1 - \phi_{Y_p}^*(M_{\delta(Y_p)}) \right)^2 \tau^2 \right].
\end{aligned} \tag{18}$$

The first inequality holds by the second part of Assumption 4 and $N_s^{\delta(Y_p)} \leq M_{\delta(Y_p)}, \forall s \in Y_p$. We let $\phi_{Y_p}^{**}(M_{\delta(Y_p)}) := M_{\delta(Y_p)} |Y_p| \tau^2 / (M_{\delta(Y_p)} |Y_p| \tau^2 + \sigma_{\min}^2)$ which minimizes

$$\frac{\phi_{Y_p}(M_{\delta(Y_p)})^2}{M_{\delta(Y_p)}} \sigma_{\min}^2 + \sum_{s \in Y_p} (1 - \phi_{Y_p}(M_{\delta(Y_p)}))^2 \tau^2,$$

for any realization of $\delta(Y_p)$. This yields,

$$(18) \geq \mathbb{E} \left[\frac{\phi_{Y_p}^{**}(M_{\delta(Y_p)})^2}{M_{\delta(Y_p)}} \right] \sigma_{\min}^2 + \sum_{s \in Y_p} \mathbb{E} \left[\left(1 - \phi_{Y_p}^{**}(M_{\delta(Y_p)}) \right)^2 \tau^2 \right].$$

We now consider two scenarios. First, when $M_{\delta(Y_p)} \geq 1$,

$$\begin{aligned}
&\mathbb{E} \left[\frac{\phi_{Y_p}^{**}(M_{\delta(Y_p)})^2}{M_{\delta(Y_p)}} \right] \sigma_{\min}^2 + \sum_{s \in Y_p} \mathbb{E} \left[\left(1 - \phi_{Y_p}^{**}(M_{\delta(Y_p)}) \right)^2 \tau^2 \right] \\
&\geq \mathbb{E} \left[\frac{\phi_{Y_p}^{**}(M_{\delta(Y_p)})^2}{M_{\delta(Y_p)}} \right] \sigma_{\min}^2 \\
&\geq \phi_{Y_p}^{**}(1) \frac{\phi_{Y_p}^{**}(\mathbb{E}[M_{\delta(Y_p)}])}{\mathbb{E}[M_{\delta(Y_p)}]} \sigma_{\min}^2 \\
&= \frac{|Y_p| \tau^2}{|Y_p| \tau^2 + \sigma_{\min}^2} \cdot \frac{|Y_p| \tau^2}{\mathbb{E}[M_{\delta(Y_p)}] |Y_p| \tau^2 + \sigma_{\min}^2} \cdot \sigma_{\min}^2 \\
&\geq \frac{\tau^2}{\tau^2 + \sigma_{\min}^2} \cdot \frac{\tau^2}{\mathbb{E}[M_{\delta(Y_p)}] \tau^2 + \sigma_{\min}^2} \cdot \sigma_{\min}^2.
\end{aligned}$$

The second inequality holds because $\phi_{y_p}^{**}(\cdot)$ is non-decreasing so that $\phi_{y_p}^{**}(M_{\delta(Y_p)}) \geq \phi_{y_p}^{**}(1)$ and $\mathbb{E}[\phi_{Y_p}^{**}(M_{\delta(Y_p)})/M_{\delta(Y_p)}] \geq \phi_{Y_p}^{**}(\mathbb{E}[M_{\delta(Y_p)}])/\mathbb{E}[M_{\delta(Y_p)}]$ by Jensen's inequality because the term $\phi_{Y_p}^{**}(M_{\delta(Y_p)})/M_{\delta(Y_p)}$ is convex in $M_{\delta(Y_p)}$. The last inequality holds as $|Y_p| \geq 1$. Note that the final term $(\tau^2/(\tau^2 + \sigma_{\min}^2)) \cdot (\tau^2/(\mathbb{E}[M_{\delta(Y_p)}] \tau^2 + \sigma_{\min}^2)) \cdot \sigma_{\min}^2 \leq \tau^2$.

On the other hand, when $M_{\delta(Y_p)} = 0$,

$$\mathbb{E} \left[\frac{\phi_{Y_p}^{**}(M_{\delta(Y_p)})^2}{M_{\delta(Y_p)}} \right] \sigma_{\min}^2 + \sum_{s \in Y_p} \mathbb{E} \left[\left(1 - \phi_{Y_p}^{**}(M_{\delta(Y_p)}) \right)^2 \tau^2 \right] = |Y_p| \tau^2 \geq \tau^2.$$

This suggests that for all Y_p ,

$$(18) \geq \frac{\tau^2}{\tau^2 + \sigma_{\min}^2} \cdot \frac{\tau^2}{\mathbb{E}[M_{\delta(Y_p)}] \tau^2 + \sigma_{\min}^2} \cdot \sigma_{\min}^2.$$

By taking the expectation over Y_p ,

$$R \left(\hat{\Theta}_{Y_p}^{*(\text{route})} \right) \geq \frac{\tau^2}{\tau^2 + \sigma_{\min}^2} \cdot \frac{\tau^2}{Nq_{\delta}\tau^2 + \sigma_{\min}^2} \cdot \sigma_{\min}^2 = \Omega(1/(Nq_{\delta})). \quad (19)$$

Based on (17) and (19), we have that when $q_{\delta} = o(1/|\mathcal{S}_p|)$,

$$\lim_{p \rightarrow \infty} \frac{R \left(\hat{\Theta}_{Y_p}^{(\text{seg})} \right)}{R \left(\hat{\Theta}_{Y_p}^{*(\text{route})} \right)} = 0.$$

This completes the proof. □

Proof of Lemma 2. We first derive the lower bounds. For even p , the probability mass function (PMF) of the symmetric beta-binomial distribution is symmetric about $p/2$, and has a minimum value on its support at $p/2$. For simplicity and without loss of generality, we will assume p is even in this proof. The odd case can be proven with some minor modifications. We have,

$$\begin{aligned} \mathbb{P}[x = (p/2, \cdot)] &= \binom{p}{p/2} \frac{B(\alpha + p/2, \alpha + p/2)}{B(\alpha, \alpha)} \\ &= \frac{1}{B(\alpha, \alpha)} \frac{\Gamma(p+1)}{\Gamma(p/2+1)\Gamma(p/2+1)} \cdot \frac{\Gamma(p/2+\alpha)\Gamma(p/2+\alpha)}{\Gamma(p+2\alpha)}, \end{aligned}$$

where $B(\cdot, \cdot)$ is Beta function and $\Gamma(\cdot)$ is Gamma function.

Define by

$$f(p) := \frac{\Gamma(p+1)}{\Gamma(p/2+1)\Gamma(p/2+1)} \frac{\Gamma(p/2+\alpha)\Gamma(p/2+\alpha)}{\Gamma(p+2\alpha)}.$$

By Gautschi's inequality [DLMF, Eq. 5.6.4], we have,

$$\begin{aligned} x^{1-\beta} &\leq \frac{\Gamma(x+1)}{\Gamma(x+\beta)} \leq (x+1)^{1-\beta}, \quad 0 < \beta \leq 1; \\ (x+1)^{1-\beta} &\leq \frac{\Gamma(x+1)}{\Gamma(x+\beta)} \leq x^{1-\beta}, \quad 1 < \beta \leq 2. \end{aligned}$$

- Under the case that $0 < \alpha \leq 1/2$, it follows that

$$\begin{aligned}\frac{\Gamma(p+1)}{\Gamma(p+2\alpha)} &\geq p^{1-2\alpha}, \\ \frac{\Gamma(p/2+1)}{\Gamma(p/2+\alpha)} &\leq (p/2+1)^{1-\alpha},\end{aligned}$$

and so

$$\frac{\Gamma(p/2+\alpha)}{\Gamma(p/2+1)} \geq (p/2+1)^{\alpha-1}.$$

It follows that

$$\begin{aligned}f(p) &\geq (p/2+1)^{2\alpha-2} p^{1-2\alpha} \\ &= 2^{2-2\alpha} \frac{p}{(p+2)^2} \left(\frac{p+2}{p}\right)^{2\alpha} \\ &\geq 2^{2-2\alpha} \frac{p}{(p+2)^2} \\ &\geq \frac{1}{9} 2^{2-2\alpha} p^{-1},\end{aligned}$$

and therefore

$$\mathbb{P}[x = (i, \cdot)] \geq \mathbb{P}[x = (p/2, \cdot)] \geq \frac{4^{1-\alpha}}{9B(\alpha, \alpha)} p^{-1}.$$

This gives,

$$\mathbb{P}[x = (i, j)] \geq \mathbb{P}[x = (p/2, p/2)] \geq \frac{4^{2-2\alpha}}{81B(\alpha, \alpha)^2} p^{-2}.$$

- Under the case that $1/2 < \alpha \leq 1$, it follows that,

$$\begin{aligned}\frac{\Gamma(p+1)}{\Gamma(p+2\alpha)} &\geq (p+1)^{1-2\alpha}, \\ \frac{\Gamma(p/2+1)}{\Gamma(p/2+\alpha)} &\leq (p/2+1)^{1-\alpha},\end{aligned}$$

and so

$$\frac{\Gamma(p/2+\alpha)}{\Gamma(p/2+1)} \geq (p/2+1)^{\alpha-1}.$$

It follows that

$$\begin{aligned}f(p) &\geq (p/2+1)^{2\alpha-2} (p+1)^{1-2\alpha} \\ &= 2^{2-2\alpha} (p+2)^{2\alpha-2} (p+1)^{1-2\alpha} \\ &= 2^{2-2\alpha} \frac{(p+2)^{2\alpha-2}}{(p+1)^{2\alpha-1}}\end{aligned}$$

$$\begin{aligned}
&= 2^{2-2\alpha} \frac{(p+2)^{2\alpha-2}}{(p+1)^{2\alpha-2}} \frac{1}{p+1} \\
&> 2^{2-2\alpha} \frac{1}{p+1} \\
&\geq 2^{2-2\alpha} \frac{1}{2p} = 2^{1-2\alpha} p^{-1},
\end{aligned}$$

and therefore

$$\mathbb{P}[x = (i, \cdot)] \geq \mathbb{P}[x = (p/2, \cdot)] \geq \frac{2^{1-2\alpha}}{B(\alpha, \alpha)} p^{-1}.$$

This gives,

$$\mathbb{P}[x = (i, j)] \geq \mathbb{P}[x = (p/2, p/2)] \geq \frac{4^{1-2\alpha}}{B(\alpha, \alpha)^2} p^{-2}.$$

We note that using the other side of Gautschi's inequality, we can also show that $\mathbb{P}[x = (p/2, p/2)] \lesssim p^{-2}$. This gives $\mathbb{P}[x = (p/2, p/2)] \simeq p^{-2}$.

We then derive the upper bounds. The PMF of the symmetric beta-binomial distribution has a maximum value on its support at either 0 or p . Without loss of generality, we select the maximum at 0.

$$\begin{aligned}
\mathbb{P}[x = (0, \cdot)] &= \binom{p}{0} \frac{B(p+\alpha, \alpha)}{B(\alpha, \alpha)} \\
&= \frac{\Gamma(\alpha)}{B(\alpha, \alpha)} \cdot \frac{\Gamma(p+\alpha)}{\Gamma(p+2\alpha)}.
\end{aligned}$$

Similarly, we now look at two cases.

- Under the case that $0 < \alpha \leq 1/2$, we have

$$\begin{aligned}
\frac{\Gamma(p+\alpha)}{\Gamma(p+2\alpha)} &= \frac{\Gamma(p+\alpha)}{\Gamma(p+1)} \cdot \frac{\Gamma(p+1)}{\Gamma(p+2\alpha)} \\
&\leq p^{\alpha-1} (p+1)^{1-2\alpha} \\
&\leq p^{-\alpha},
\end{aligned}$$

and therefore

$$\mathbb{P}[x = (i, \cdot)] \leq \mathbb{P}[x = (0, \cdot)] \leq \frac{\Gamma(\alpha)}{B(\alpha, \alpha)} p^{-\alpha}.$$

This gives,

$$\mathbb{P}[x = (i, j)] \leq \mathbb{P}[x = (0, 0)] \leq \frac{\Gamma(\alpha)^2}{B(\alpha, \alpha)^2} p^{-2\alpha}.$$

- Under the case that $1/2 < \alpha \leq 1$, we have

$$\begin{aligned} \frac{\Gamma(p + \alpha)}{\Gamma(p + 2\alpha)} &= \frac{\Gamma(p + \alpha)}{\Gamma(p + 1)} \cdot \frac{\Gamma(p + 1)}{\Gamma(p + 2\alpha)} \\ &\leq p^{\alpha-1} p^{1-2\alpha} \\ &= p^{-\alpha}. \end{aligned}$$

Similarly, this gives,

$$\mathbb{P}[x = (i, j)] \leq \mathbb{P}[x = (0, 0)] \leq \frac{\Gamma(\alpha)^2}{B(\alpha, \alpha)^2} p^{-2\alpha}.$$

Using the other side of Gautschi's inequality, we can show that $\mathbb{P}[x = (0, 0)] \gtrsim p^{-2\alpha}$. By symmetry, we have $\mathbb{P}[x = (0, 0)] = \mathbb{P}[x = (p, 0)] = \mathbb{P}[x = (0, p)] = \mathbb{P}[x = (p, p)] \simeq p^{-2\alpha}$. This completes the proof. \square

Proof of Proposition 4. Consider a particular neighborhood near a predicting route y_p , $\delta^{\text{od}*}(y_p) = \{y \in \mathcal{Y}_p : \|x_1(y), x_1(y_p)\|_1 = 0, \|x_2(y), x_2(y_p)\|_1 = 0\}$. In words, neighborhood $\delta^{\text{od}*}(y_p)$ includes historical routes that have the same origin and destination as those of route y_p . It is clear that for any other route neighborhood $\delta^{\text{od}}(\cdot)$ with $\delta^{\text{od}}(y_p) = \{y \in \mathcal{Y}_p : \|x_1(y), x_1(y_p)\|_1 \leq c, \|x_2(y), x_2(y_p)\|_1 \leq c\}$ for some constant $c > 0$ that does not depend on p ,

$$q_{\delta^{\text{od}}(y_p)} \simeq q_{\delta^{\text{od}*}(y_p)}.$$

We thus focus on analyzing $q_{\delta^{\text{od}*}} = \mathbb{P}_{Y_p \sim \mu_p, Y'_p \sim \mu_p}[Y_p \in \delta^{\text{od}*}(Y'_p)] = \sum_{y \in \mathcal{Y}_p} q_{\delta^{\text{od}*}(y)} \mathbb{P}[Y'_p = y]$ instead.

$$\begin{aligned} q_{\delta^{\text{od}*}} &= \sum_{y \in \mathcal{Y}_p} q_{\delta^{\text{od}*}(y)} \mathbb{P}[Y'_p = y] \\ &= \sum_{x_1 \in \mathcal{V}_p, x_2 \in \mathcal{V}_p} \mathbb{P}[x_1(Y_p) = x_1, x_2(Y_p) = x_2] \cdot \mathbb{P}[x_1(Y'_p) = x_1, x_2(Y'_p) = x_2] \\ &= \sum_{x_1 \in \mathcal{V}_p, x_2 \in \mathcal{V}_p} \mathbb{P}^2[x_1(Y_p) = x_1, x_2(Y_p) = x_2] \\ &= \sum_{x_1 \in \mathcal{V}_p, x_2 \in \mathcal{V}_p} \mathbb{P}^2[x_1(Y_p) = x_1] \mathbb{P}^2[x_2(Y_p) = x_2] \\ &= \sum_{i, j, l, m \in \{0, \dots, p\}} \mathbb{P}^2[x_1(Y_p) = (i, \cdot)] \mathbb{P}^2[x_1(Y_p) = (\cdot, j)] \mathbb{P}^2[x_2(Y_p) = (l, \cdot)] \mathbb{P}^2[x_2(Y_p) = (\cdot, m)] \\ &= \left(\sum_{i \in \{0, \dots, p\}} \mathbb{P}^2[x_1(Y_p) = (i, \cdot)] \right)^4. \end{aligned} \tag{20}$$

The third equality holds because the sampling processes of origins and destinations are independent. Similarly, the fourth equality holds because the sampling processes of horizontal and vertical coordinates are independent.

For any $i \in \{0, \dots, p-1\}$,

$$\frac{\mathbb{P}[x_1(Y_p) = (i+1, \cdot)]}{\mathbb{P}[x_1(Y_p) = (i, \cdot)]} = \frac{\binom{p}{i+1} B(i+\alpha, p-i+\alpha) / B(\alpha, \alpha)}{\binom{p}{i} B(i+\alpha, p-i+\alpha) / B(\alpha, \alpha)}$$

$$\begin{aligned}
&= \frac{\binom{p}{i+1} B(i+\alpha, p-i+\alpha)}{\binom{p}{i} B(i+\alpha, p-i+\alpha)} \\
&= \frac{p-i}{i+1} \cdot \frac{\Gamma(i+1+\alpha)\Gamma(p-i-1+\alpha)/\Gamma(p+2\alpha)}{\Gamma(i+\alpha)\Gamma(p-i+\alpha)/\Gamma(p+2\alpha)} \\
&= \frac{p-i}{i+1} \cdot \frac{\Gamma(i+1+\alpha)\Gamma(p-i-1+\alpha)}{\Gamma(i+\alpha)\Gamma(p-i+\alpha)} \\
&= \frac{p-i}{i+1} \cdot (i+\alpha) \cdot \frac{1}{p-i-1+\alpha} \\
&= \frac{i+\alpha}{i+1} \cdot \frac{p-i}{p-i-1+\alpha}. \tag{21}
\end{aligned}$$

The second-to-last equation holds by using the fact that $\Gamma(z+1) = z\Gamma(z)$. Let $g(j) = \prod_{i=1}^j \left(\frac{i+\alpha}{i+1}\right) \cdot \left(\frac{p-i}{p-i-1+\alpha}\right)$. We consider the case where the grid size p is even. The case of p being odd can be proven with minor modifications. Using recursion (21),

$$\begin{aligned}
&\sum_{i \in \{0, \dots, p\}} \mathbb{P}^2[x_1(Y_p) = (i, \cdot)] \\
&\simeq \sum_{i \in \{0, \dots, p/2-1\}} \mathbb{P}^2[x_1(Y_p) = (i, \cdot)] \tag{22}
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{P}^2[x_1(Y_p) = (0, \cdot)] \left(1 + \sum_{j=1}^{p/2-1} g^2(j)\right) \\
&\simeq p^{-2\alpha} \left(1 + \sum_{j=1}^{p/2-1} g^2(j)\right). \tag{23}
\end{aligned}$$

Equation (22) holds by the symmetry of the distributions of origins and destinations. Equation (23) uses the fact that $P[x_1(Y_p) = (0, \cdot)] \simeq p^{-\alpha}$ from the proof of Lemma 2. By Lemma 4 in Appendix C which shows that $g(j) \simeq (1/(j+1))^{1-\alpha}$ for all $j \leq p/2-1$,

$$\begin{aligned}
(22) &\simeq p^{-2\alpha} \left(1 + \sum_{j=1}^{p/2-1} \left(\frac{1}{j+1}\right)^{2-2\alpha}\right) \\
&\simeq p^{-2\alpha} \left(\sum_{j=1}^{p/2} \left(\frac{1}{j}\right)^{2-2\alpha}\right)
\end{aligned}$$

- When $0 < \alpha < 1/2$, we know that $\sum_{j=1}^{\infty} (1/j)^{2-2\alpha} < +\infty$ converges. As a result, $\sum_{i \in \{0, \dots, p\}} \mathbb{P}^2[x_1(Y_p) = (i, \cdot)]$ as a function of p satisfies,

$$\sum_{i \in \{0, \dots, p\}} \mathbb{P}^2[x_1(Y_p) = (i, \cdot)] \simeq \sum_{i \in \{0, \dots, p/2-1\}} \mathbb{P}^2[x_1(Y_p) = (i, \cdot)] \simeq p^{-2\alpha}.$$

- When $1/2 \leq \alpha \leq 1$, we know that $\sum_{j=1}^{\infty} (1/j)^{2-2\alpha}$ diverges. For $\alpha > 1/2$,

$$\int_1^{p/2} \left(\frac{1}{j+1}\right)^{2-2\alpha} dj \leq \sum_{j=1}^{p/2} \left(\frac{1}{j}\right)^{2-2\alpha} \leq \int_1^{p/2} \left(\frac{1}{j}\right)^{2-2\alpha} dj$$

$$\Leftrightarrow \frac{1}{2\alpha - 1} \left(\left(\frac{p}{2} + 1 \right)^{2\alpha - 1} - 2^{2\alpha - 1} \right) \leq \sum_{j=1}^{p/2} \left(\frac{1}{j} \right)^{2-2\alpha} \leq \frac{1}{2\alpha - 1} \left(\left(\frac{p}{2} \right)^{2\alpha - 1} - 1 \right).$$

This yields,

$$\begin{aligned} & \sum_{i \in \{0, \dots, p\}} \mathbb{P}^2[x_1(Y_p) = (i, \cdot)] \\ & \simeq \sum_{i \in \{0, \dots, p/2-1\}} \mathbb{P}^2[x_1(Y_p) = (i, \cdot)] \\ & \simeq p^{-2\alpha} \left(\sum_{j=1}^{p/2} \left(\frac{1}{j} \right)^{2-2\alpha} \right) \\ & \simeq p^{-2\alpha} \cdot p^{2\alpha-1} \\ & \simeq p^{-1}. \end{aligned}$$

Plugging these rates into (20) completes the proof. \square

Proof of Proposition 5. Consider a segment $s = (i, j) \rightarrow (i+1, j) \in \mathcal{S}_p$ from the grid of even size p and assume that $i < p/2$. By symmetry, segments with vertical movements or at other positions can be proven in the same way. The case with odd p can be proven with minor modifications. Consider a route $Y_p \sim \mu_p$ that covers segment s . Let X_1 and X_2 be the corresponding origin and destination of route Y_p . There are two scenarios in which $s \in Y_p$.

- $X_1 = (i_1, j)$ for some $i_1 \leq i$ and $X_2 = (i_2, \cdot)$ for some $i_2 > i$. This is with probability

$$\begin{aligned} & \simeq \sum_{i_1=0}^i \sum_{i_2=i+1}^p \mathbb{P}(X_1 = (i_1, j)) \mathbb{P}(X_2 = (i_2, \cdot)) \\ & = \mathbb{P}(X_1 = (\cdot, j)) \cdot \left(\sum_{i_1=0}^i \sum_{i_2=i+1}^p \mathbb{P}(X_1 = (i_1, \cdot)) \mathbb{P}(X_2 = (i_2, \cdot)) \right) \\ & = \mathbb{P}(X_1 = (\cdot, j)) \cdot \left(\sum_{i_1=0}^i \mathbb{P}(X_1 = (i_1, \cdot)) \right) \left(\sum_{i_2=i+1}^p \mathbb{P}(X_2 = (i_2, \cdot)) \right). \end{aligned} \quad (24)$$

Clearly, given i , (24) achieves its minimum value when $j = p/2$ and achieves its maximum value when $j = 0$ or p . Similarly, given j , we can show that (24) achieves its minimum value at $i = 0$ and its maximum value at $i = p/2 - 1$. To see that, for any $i < p/2$,

$$\sum_{i_1=0}^i \mathbb{P}(X_1 = (i_1, \cdot)) < \sum_{i_2=i+1}^p \mathbb{P}(X_2 = (i_2, \cdot)).$$

This yields,

$$\left(\sum_{i_1=0}^{i-1} \mathbb{P}(X_1 = (i_1, \cdot)) \right) \left(\sum_{i_2=i}^p \mathbb{P}(X_2 = (i_2, \cdot)) \right)$$

$$\begin{aligned}
&= \left(\sum_{i_1=0}^i \mathbb{P}(X_1 = (i_1, \cdot)) - \mathbb{P}(X_1 = (i, \cdot)) \right) \left(\sum_{i_2=i+1}^p \mathbb{P}(X_2 = (i_2, \cdot)) + \mathbb{P}(X_2 = (i, \cdot)) \right) \\
&= \left(\sum_{i_1=0}^i \mathbb{P}(X_1 = (i_1, \cdot)) - \mathbb{P}(X_1 = (i, \cdot)) \right) \left(\sum_{i_2=i+1}^p \mathbb{P}(X_2 = (i_2, \cdot)) + \mathbb{P}(X_1 = (i, \cdot)) \right) \\
&= \left(\sum_{i_1=0}^i \mathbb{P}(X_1 = (i_1, \cdot)) \right) \left(\sum_{i_2=i+1}^p \mathbb{P}(X_2 = (i_2, \cdot)) \right) \\
&\quad - \mathbb{P}(X_1 = (i, \cdot)) \left(\sum_{i_2=i+1}^p \mathbb{P}(X_2 = (i_2, \cdot)) - \sum_{i_1=0}^i \mathbb{P}(X_1 = (i_1, \cdot)) \right) \\
&\quad - \mathbb{P}^2(X_1 = (i, \cdot)) \\
&\leq \left(\sum_{i_1=0}^i \mathbb{P}(X_1 = (i_1, \cdot)) \right) \left(\sum_{i_2=i+1}^p \mathbb{P}(X_2 = (i_2, \cdot)) \right),
\end{aligned}$$

for all $i < p/2$. This suggests that (24) achieves its overall maximum at $i = p/2 - 1, j = 0$ or p with

$$\sum_{i_1=0}^{p/2-1} \sum_{i_2=p/2}^p \mathbb{P}(X_1 = (i_1, 0)) \mathbb{P}(X_2 = (i_2, \cdot)) \simeq \mathbb{P}(X_1 = (\cdot, 0)) \simeq p^{-\alpha}.$$

On the other hand, it achieves its overall minimum at $i = 0, j = p/2$ with

$$\sum_{i_1=0}^0 \sum_{i_2=1}^p \mathbb{P}(X_1 = (i_1, p/2)) \mathbb{P}(X_2 = (i_2, \cdot)) \simeq \mathbb{P}(X_1 = (0, p/2)) \simeq p^{-1-\alpha}.$$

- $X_1 = (i_1, \cdot)$ for some $i_1 \leq i$ and $x_2 = (i_2, j)$ for some $i_2 > i$. This is with probability

$$\begin{aligned}
&\simeq \sum_{i_1=0}^i \sum_{i_2=i+1}^p \mathbb{P}(X_1 = (i_1, \cdot)) \mathbb{P}(X_2 = (i_2, j)) \\
&= \mathbb{P}(X_2 = (\cdot, j)) \cdot \left(\sum_{i_1=0}^i \sum_{i_2=i+1}^p \mathbb{P}(X_1 = (i_1, \cdot)) \mathbb{P}(X_2 = (i_2, \cdot)) \right) \\
&= \mathbb{P}(X_2 = (\cdot, j)) \cdot \left(\sum_{i_1=0}^i \mathbb{P}(X_1 = (i_1, \cdot)) \right) \left(\sum_{i_2=i+1}^p \mathbb{P}(X_2 = (i_2, \cdot)) \right),
\end{aligned}$$

which is symmetric to the previous case and thus has the same conclusion.

This concludes the proof. □

Proof of Theorem 4. We first give a lower bound for $R(\hat{\Theta}_y^*)$. By Lemma 1,

$$R\left(\hat{\Theta}_y^* \middle| y_{[N]}\right) \geq \frac{|y|^2}{\sum_{s,t \in y} N_{s \cup t} \psi_{s,t} + |y|/\tau^2}. \tag{25}$$

Under a given road network size p , let indicator random variable I_s denote whether a road segment s is traversed by a randomly sampled route $Y_p \sim \mu_p$. Given $\{I_s\}_{s \in \mathcal{S}_p}$, we rewrote (25) as

$$\begin{aligned} R\left(\hat{\Theta}_y^* \middle| y_{[N]}\right) &\geq \frac{\left(\sum_{s \in \mathcal{S}_p} I_s\right)^2}{\sum_{s,t \in \mathcal{S}_p} N_{s \cup t} \psi_{s,t} I_s I_t + \left(\sum_{s \in \mathcal{S}_p} I_s\right) / \tau^2} \\ &= \frac{1}{\sum_{s,t \in \mathcal{S}_p} N_{s \cup t} \psi_{s,t} \frac{I_s I_t}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} + \frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right) \tau^2}} \\ &\geq \frac{1}{\sum_{s,t \in \mathcal{S}_p} N_{s \cup t} |\psi_{s,t}| \frac{I_s I_t}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} + \frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right) \tau^2}}. \end{aligned}$$

Taking expectations over the predicting route and the historical routes,

$$\begin{aligned} R(\hat{\Theta}_{Y_p}^*) &\geq \mathbb{E} \left[\frac{1}{\sum_{s,t \in \mathcal{S}_p} N_{s \cup t} |\psi_{s,t}| \frac{I_s I_t}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} + \frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right) \tau^2}} \right] \\ &\geq \frac{1}{\sum_{s,t \in \mathcal{S}_p} \mathbb{E}[N_{s \cup t}] |\psi_{s,t}| \mathbb{E} \left[\frac{I_s I_t}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} \right] + \frac{1}{\tau^2} \mathbb{E} \left[\frac{1}{\sum_{s' \in \mathcal{S}_p} I_{s'}} \right]}. \end{aligned}$$

The second inequality holds because of $\mathbb{E}[1/X] \geq 1/\mathbb{E}[X]$ for any non-negative random variable X . By Theorem 3, we have $R(\hat{\Theta}_{Y_p}^{(\text{seg})}) = \mathcal{O}(p^2/N)$. This gives,

$$\begin{aligned} \frac{R(\hat{\Theta}_{Y_p}^{(\text{seg})})}{R(\hat{\Theta}_{Y_p}^*)} &\leq \mathcal{O}(p^2/N) \left(\sum_{s,t \in \mathcal{S}_p} \mathbb{E}[N_{s \cup t}] |\psi_{s,t}| \mathbb{E} \left[\frac{I_s I_t}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} \right] \right. \\ &\quad \left. + \frac{1}{\tau^2} \mathbb{E} \left[\frac{1}{\sum_{s' \in \mathcal{S}_p} I_{s'}} \right] \right). \end{aligned} \tag{26}$$

We analyze the two terms in the parenthesis separately. For the first term,

$$\begin{aligned} &\mathcal{O}(p^2/N) \sum_{s,t \in \mathcal{S}_p} \mathbb{E}[N_{s \cup t}] |\psi_{s,t}| \mathbb{E} \left[\frac{I_s I_t}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} \right] \\ &= \mathcal{O}(p^2) \sum_{s,t \in \mathcal{S}_p} \frac{\mathbb{E}[N_{s \cup t}]}{N} |\psi_{s,t}| \mathbb{E} \left[\frac{I_s I_t}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} \right] \\ &\leq \mathcal{O}(p^2) \sum_{s,t \in \mathcal{S}_p} \mathbb{P}(I_s = 1, I_t = 1) |\psi_{s,t}| \mathbb{E} \left[\frac{I_s}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} \right] \\ &\leq \mathcal{O}(p^2) \sum_{s,t \in \mathcal{S}_p} \mathbb{P}(I_s = 1) |\psi_{s,t}| \mathbb{E} \left[\frac{I_s}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} \right] \\ &= \mathcal{O}(p^2) \sum_{s \in \mathcal{S}_p} \sum_{t \in \mathcal{S}_p} \mathbb{P}^2(I_s = 1) |\psi_{s,t}| \mathbb{E} \left[\frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} \middle| I_s = 1 \right] \\ &= \mathcal{O}(p^2) \sum_{s \in \mathcal{S}_p} \mathbb{P}^2(I_s = 1) \mathbb{E} \left[\frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'}\right)^2} \middle| I_s = 1 \right] \left(\sum_{t \in \mathcal{S}_p} |\psi_{s,t}| \right) \end{aligned}$$

$$\begin{aligned}
&= \mathcal{O}(p^2) \sum_{s \in \mathcal{S}_p} \mathbb{P}^2(I_s = 1) \mathbb{E} \left[\frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'} \right)^2} \middle| I_s = 1 \right] \mathcal{O}(1) \\
&= \mathcal{O}(p^2) \sum_{s \in \mathcal{S}_p} \mathbb{P}^2(I_s = 1) \mathbb{E} \left[\frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'} \right)^2} \middle| I_s = 1 \right]. \tag{27}
\end{aligned}$$

The second-to-last equality uses the second part of Assumption 4. We have the following claims whose proof can be found in the proof of Lemma 7 in Appendix C. For any segment $s \in \mathcal{S}_p$,

$$\mathbb{E} \left[\frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'} \right)^2} \middle| I_s = 1 \right] = \mathcal{O}(\log(p)p^{-2}), \quad \mathbb{E} \left[\frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'} \right)^2} \middle| I_s = 0 \right] = \mathcal{O}(\log(p)p^{-2}).$$

Now go back to equation (27), for any $1/2 \leq \alpha \leq 1$,

$$\begin{aligned}
&\mathcal{O}(p^2) \sum_{s \in \mathcal{S}_p} \mathbb{P}^2(I_s = 1) \mathbb{E} \left[\frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'} \right)^2} \middle| I_s = 1 \right] \\
&= \mathcal{O}(p^2) \left(\sum_{s \in \mathcal{S}_p} \mathbb{P}^2(I_s = 1) \right) \mathcal{O}(\log(p)p^{-2}) \\
&= \mathcal{O}(p^2) \mathcal{O}(1) \mathcal{O}(\log(p)p^{-2}) \quad (\text{by Lemma 5 in Appendix C}) \\
&= \mathcal{O}(\log(p)).
\end{aligned}$$

For the second term in (26),

$$\begin{aligned}
&\mathcal{O}(p^2/N) \frac{1}{\tau^2} \mathbb{E} \left[\frac{1}{\sum_{s' \in \mathcal{S}_p} I_{s'}} \right] \\
&\leq \mathcal{O}(p^2/N) \frac{1}{\tau^2} \sqrt{\mathbb{E} \left[\frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'} \right)^2} \right]} \\
&= \mathcal{O}(p^2/N) \frac{1}{\tau^2} \sqrt{\mathbb{E} \left[\frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'} \right)^2} \middle| I_s = 1 \right] \mathbb{P}(I_s = 1) + \mathbb{E} \left[\frac{1}{\left(\sum_{s' \in \mathcal{S}_p} I_{s'} \right)^2} \middle| I_s = 0 \right] \mathbb{P}(I_s = 0)} \\
&= \mathcal{O}(p^2/N) \frac{1}{\tau^2} \sqrt{\mathcal{O}(p^{-2} \log(p))} \\
&= \mathcal{O} \left(p \sqrt{\log(p)} / N \right) \frac{1}{\tau^2} \\
&= \mathcal{O}(1). \quad (\text{as } N = \omega(p))
\end{aligned}$$

The first inequality uses the fact that $\mathbb{E}[X] \leq \sqrt{\mathbb{E}[X^2]}$ for any random variable X .

This completes the proof that $R(\hat{\Theta}_{Y_p}^{(\text{seg})})/R(\hat{\Theta}_{Y_p}^*) = \mathcal{O}(\log(p))$ when $N = \omega(p)$ and $1/2 \leq \alpha \leq 1$. For the second part of the theorem, using the information-theoretic lower bound (Lemma 1), given

a set of historical routes $Y_{p,[N]}$ and the predicting route Y_p under a grid size p , for any estimator,

$$\begin{aligned}
R\left(\hat{\Theta}_{Y_p} \mid Y_{p,[N]}\right) &\geq \frac{|Y_p|^2}{\sum_{s,t \in Y_p} N_{s \cup t} \psi_{s,t} + |Y_p|/\tau^2} \\
&\geq \frac{|Y_p|^2}{\sum_{s,t \in Y_p} N \psi_{s,t} + |Y_p|/\tau^2} && (N \geq N_{s \cup t}, \forall s, t \in Y_p) \\
&= \frac{|Y_p|^2}{\sum_{s,t \in Y_p} \mathcal{O}(p) \psi_{s,t} + |Y_p|/\tau^2} \\
&= \frac{|Y_p|^2}{|Y_p| \mathcal{O}(p) + |Y_p|/\tau^2} && (\sum_{t \in Y_p} \psi_{s,t} = \mathcal{O}(1) \text{ by Assumption 2.2}) \\
&= \frac{|Y_p|}{\mathcal{O}(p) + 1/\tau^2}.
\end{aligned}$$

Taking the expectation over the predicting route $Y_p \in \mu_p$ yields

$$R\left(\hat{\Theta}_{Y_p}\right) \geq \frac{\mathbb{E}[|Y_p|]}{\mathcal{O}(p) + 1/\tau^2}.$$

We know that as α increases, route origins and destinations are distributed toward the center of the grid. We thus can focus on the case of $\alpha = 1$ to get a lower bound for $\mathbb{E}[|Y_p|]$. When $\alpha = 1$,

$$\begin{aligned}
\mathbb{E}[|Y_p|] &= \frac{1}{(p+1)^2} \left(\sum_{i_1=0}^p \sum_{i_2=0}^p |i_1 - i_2| + \sum_{j_1=0}^p \sum_{j_2=0}^p |j_1 - j_2| \right) \\
&= \frac{2}{(p+1)^2} \left(\sum_{i_1=0}^p \sum_{i_2=0}^p |i_1 - i_2| \right) \\
&= \frac{2}{(p+1)^2} \left(\sum_{i_1=0}^p \sum_{i_2=i_1+1}^p (i_2 - i_1) + \sum_{i_1=0}^p \sum_{i_2=0}^{i_1-1} (i_1 - i_2) \right) \\
&= \frac{2}{(p+1)^2} \cdot 2 \cdot \sum_{i=1}^p \frac{i(i+1)}{2} \\
&= \frac{2}{(p+1)^2} \left(\sum_{i=1}^p i^2 + \sum_{i=1}^p i \right) \\
&= \frac{2}{(p+1)^2} \left(\frac{p(p+1)(2p+1)}{6} + \frac{p(p+1)}{2} \right) \\
&= \frac{p(2p+1)}{3(p+1)} + \frac{p}{p+1} \\
&= \Omega(p).
\end{aligned}$$

Thus for any $\alpha \in (0, 1]$, there exist $\epsilon > 0$ such that for any estimator,

$$R\left(\hat{\Theta}_{Y_p}\right) \geq \frac{\Omega(p)}{\mathcal{O}(p) + 1/\tau^2} > \epsilon, \quad \forall p.$$

This completes the proof that $\liminf_{p \rightarrow \infty} R\left(\hat{\Theta}_{Y_p}\right) > 0$ for any estimator. \square

Proof of Lemma 1. The result follows by adapting Theorem 1 of Gill and Levit [1995] which gives a multivariate version of the van Trees inequality. Consider estimating the total travel time on route y , $\Theta_y = \sum_{s \in y} \theta_s$, with one single observation $z \in \mathbb{R}_{\geq 0}^{M \times 1}$ where z_i is a single observed travel time for a single segment on a single trip and $M = \sum_{n=1}^N |y_n|$. Let $u_i = s$ if the i^{th} observation in Z is a travel time on segment s . Let $w_i = n$ if the i^{th} observation in z is a travel time from trip n . Thus, $z_i = T'_{w_i, u_i}$, $\forall i \in \{1, \dots, M\}$. With a bit abuse of notation, let $\theta_y = [\theta_s]_{s \in y}$. Define the Fisher information matrix

$$\mathcal{I}(\theta_y) = \mathbb{E} \left[\left(\frac{\partial \log f(Z | \theta_y)}{\partial \theta_y} \right)^\top \left(\frac{\partial \log f(Z | \theta_y)}{\partial \theta_y} \right) \right] \in \mathbb{R}^{|y| \times |y|},$$

where the expectation is taken over Z with u_i and w_i fixed and $f(Z | \theta_y)$ is the density of Z given θ_y . Suppose $f(\cdot | \theta_y)$ is on an arbitrary measure space \mathbf{Z} for all θ_y . Note that $\partial \log f(Z | \theta_y) / \partial \theta_y := [\partial \log f(Z | \theta_y) / \partial \theta_s]_{s \in y} \in \mathbb{R}^{1 \times |y|}$. Let $\lambda(\theta_y)$ be the prior density. Suppose $\theta_y \in \Theta_y \in \mathbb{R}^{1 \times |y|}$ and $\theta_s \in \Theta_s \in \mathbb{R}$. Define the information on the prior distribution $\lambda(\cdot)$,

$$\mathcal{I}(\lambda) = \mathbb{E} \left[\left(\frac{\partial \log \lambda(\theta_y)}{\partial \theta_y} \right)^\top \left(\frac{\partial \log \lambda(\theta_y)}{\partial \theta_y} \right) \right] \in \mathbb{R}^{|y| \times |y|},$$

where the expectation is taken over θ_y . The following result is taken from Theorem 1 of Gill and Levit [1995] and adapted to our setup by choosing $B(\theta_y) = 1$ and $C(\theta_y) = 1^{1 \times |y|}$ in their theorem. The original assumptions stated in Gill and Levit [1995] are provided below. We call a function $g(\theta_y)$ nice if for each $s \in y$, it is absolutely continuous in θ_s for almost all values of the other components of θ_y and its partial derivatives $\partial g / \partial \theta_s$ are measurable in θ_y .

Assumption 5. Gill and Levit [1995] impose the following assumptions.

1. $f(z | \theta_y)$ is nice in θ_y for almost all z and its partial derivatives with respect to θ_y are measurable in z, θ_y .
2. The Fisher information matrix $\mathcal{I}(\theta_y)$ exists and $\text{diag}(\mathcal{I}(\theta_y))^{1/2}$ is locally integrable in θ_y .
3. $\lambda(\theta_y)$ is nice in θ_y ; Θ_y is compact with boundary which is piecewise C^1 -smooth; $\lambda(\theta_y)$ is positive on the interior of Θ_y and zero on its boundary.

Theorem 6 (Multivariate van Trees inequality). *For any estimator $\hat{\Theta}_y$,*

$$\begin{aligned} & \int_{\Theta_y} \mathbb{E} \left[\left(\hat{\Theta}_y - \Theta_y \right)^2 \middle| y_{[N]}, \theta_y \right] \lambda(\theta_y) d\theta_y \\ & \geq \frac{|y|^2}{\int_{\Theta_y} \text{trace}(\mathcal{I}(\theta_y)) \lambda(\theta_y) d\theta_y + \text{trace}(\mathcal{I}(\lambda))}. \end{aligned} \quad (28)$$

Revised the third part of Assumption 5. We note that the compactness of Θ_y can be replaced with $\Theta_y = \mathbb{R}^{1 \times |y|}$ and $\lim_{\theta_s \rightarrow +\infty} \Theta_y \lambda(\theta_y) = 0$ and $\lim_{\theta_s \rightarrow -\infty} \Theta_y \lambda(\theta_y) = 0$ for all $s \in y$ and for almost all values of the other components of θ_y . We still require $\lambda(\theta_y)$ to be nice in $\theta_y \in \Theta_y$. We will use this changed assumption later in the proof as we will impose a Gaussian prior for $\lambda(\theta_y)$. Here we provide an updated proof based on this revised assumption.

Proof of Theorem 6 under revised Assumption 5. Most of the proof follows exactly as the one provided on page 65 of Gill and Levit [1995]. We do not repeat the arguments here. The only part we have to re-verify is the derivation of $\mathbb{E}[XY]$ where $X = \hat{\Theta}_y - \Theta_y$ and $Y = \sum_{s \in y} (\partial \{f(Z | \theta_y) \lambda(\theta_y)\} / \partial \theta_s)$.

$(1/(f(Z | \theta_y)\lambda(\theta_y)))$. We let Θ_{-s} and θ_{-s} define the measure space and vector excluding the s^{th} component.

$$\begin{aligned}
\mathbb{E}[XY] &= \int_{\mathbf{Z}} \int_{\Theta_y} (\hat{\Theta}_y - \Theta_y) \sum_{s \in y} \frac{\partial \{f(z | \theta_y)\lambda(\theta_y)\}}{\partial \theta_s} d\theta dz \\
&= \int_{\mathbf{Z}} \left(\sum_{s \in y} \int_{\Theta_{-s}} \int_{\Theta_s} (\hat{\Theta}_y - \Theta_y) \frac{\partial \{f(z | \theta_y)\lambda(\theta_y)\}}{\partial \theta_s} d\theta_s d\theta_{-s} \right) dz \\
&= \int_{\mathbf{Z}} \left(\sum_{s \in y} \int_{\Theta_{-s}} \int_{-\infty}^{+\infty} (\hat{\Theta}_y - \Theta_y) \frac{\partial \{f(z | \theta_y)\lambda(\theta_y)\}}{\partial \theta_s} d\theta_s d\theta_{-s} \right) dz \\
&= \int_{\mathbf{Z}} \left(\sum_{s \in y} \int_{\Theta_{-s}} \left(\left[(\hat{\Theta}_y - \Theta_y) f(z | \theta_y)\lambda(\theta_y) \right]_{-\infty}^{+\infty} + \int_{-\infty}^{+\infty} f(z | \theta_y)\lambda(\theta_y) d\theta_s \right) d\theta_{-s} \right) dz \\
&= \int_{\mathbf{Z}} \left(\sum_{s \in y} \int_{\Theta_{-s}} \left(\int_{-\infty}^{+\infty} f(z | \theta_y)\lambda(\theta_y) d\theta_s \right) d\theta_{-s} \right) dz \\
&= |y| \int_{\mathbf{Z}} \int_{\Theta_y} f(z | \theta_y)\lambda(\theta_y) d\theta_y dz \\
&= |y|.
\end{aligned}$$

The fourth equality is integration by parts. The fifth equality holds as $\lim_{\theta_s \rightarrow +\infty} \Theta_y \lambda(\theta_y) = 0$ and $\lim_{\theta_s \rightarrow -\infty} \Theta_y \lambda(\theta_y) = 0$, which implies that $\lim_{\theta_s \rightarrow +\infty} \lambda(\theta_y) = 0$ and $\lim_{\theta_s \rightarrow -\infty} \lambda(\theta_y) = 0$. The rest of the proof follows exactly as the one in Gill and Levit [1995] by showing that

$$\mathbb{E}[Y^2] = \int_{\Theta_y} \text{trace}(\mathcal{I}(\theta_y))\lambda(\theta_y) d\theta_y + \text{trace}(\mathcal{I}(\lambda)),$$

and finally by Cauchy-Schwarz inequality,

$$\begin{aligned}
\int_{\Theta_y} \mathbb{E} \left[\left(\hat{\Theta}_y - \Theta_y \right)^2 \middle| y_{[N]}, \theta_y \right] \lambda(\theta_y) d\theta_y &= \mathbb{E}[X^2] \geq \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]} \\
&= \frac{|y|^2}{\int_{\Theta_y} \text{trace}(\mathcal{I}(\theta_y))\lambda(\theta_y) d\theta_y + \text{trace}(\mathcal{I}(\lambda))}.
\end{aligned}$$

This completes the proof. \square

By the data generative process, we know that

$$f(z | \theta_y) = \prod_{n=1}^N f_n(\{z_i : w_i = n\} | \theta_y) \Rightarrow \log f(z | \theta_y) = \sum_{n=1}^N \log f_n(\{z_i : w_i = n\} | \theta_y),$$

where $f_n(\{z_i : w_i = n\} | \theta_y)$ is the density of observing the segment travel times on trip n given θ_y . Define the Fisher information matrix for each trip n ,

$$\mathcal{I}_n(\theta_y) = \mathbb{E} \left[\left(\frac{\partial \log f_n(\{Z_i : w_i = n\} | \theta_y)}{\partial \theta_y} \right)^\top \left(\frac{\partial \log f_n(\{Z_i : w_i = n\} | \theta_y)}{\partial \theta_y} \right) \right] \in \mathbb{R}^{|y| \times |y|}.$$

By an equivalent definition of the Fisher information matrix under mild regularity conditions (see Lemma 5.3 of Lehmann and Casella [2006]), we have

$$\begin{aligned}
[\mathcal{I}(\theta_y)]_{s,t} &= -\mathbb{E} \left[\frac{\partial^2}{\partial \theta_s \partial \theta_t} \log f(Z | \theta_y) \right] \\
&= -\sum_{n=1}^N \mathbb{E} \left[\frac{\partial^2}{\partial \theta_s \partial \theta_t} \log f_n(\{Z_i : w_i = n\} | \theta_y) \right] \\
&= \sum_{n=1}^N [\mathcal{I}_n(\theta_y)]_{s,t},
\end{aligned}$$

for all $s, t \in y$. Plugging this into (28) yields,

$$\begin{aligned}
&\int_{\Theta_y} \mathbb{E} \left[\left(\hat{\Theta}_y - \Theta_y \right)^2 \middle| y_{[N]}, \theta_y \right] \lambda(\theta_y) d\theta_y \\
&\geq R \left(\hat{\Theta}_y^* \middle| y_{[N]} \right) \\
&\geq \frac{|y|^2}{\sum_{n=1}^N \int_{\Theta_y} \text{tr } \mathcal{I}_n(\theta_y) \lambda(\theta_y) d\theta_y + \text{tr } \mathcal{I}(\lambda)}. \tag{29}
\end{aligned}$$

Under the assumption that (\mathcal{E}, θ) are jointly Gaussian distributed, both the means of segment travel times, and the segment travel times conditional on their means are normally distributed. Note that Gaussian priors and posteriors satisfy the revised Assumption 5. This greatly simplifies the analysis — we know that under multivariate normal the Fisher information matrix is simply the precision matrix. This gives $\mathcal{I}(\lambda) = \text{diag}(\underbrace{[1/\tau^2, \dots, 1/\tau^2]}_{|y|})$. Moreover,

$$[\mathcal{I}_n(\theta_y)]_{s,t} = \begin{cases} \psi_{s,t}, & \text{if } s, t \in y_n, \\ 0, & \text{o/w,} \end{cases}$$

for all $s, t \in y$. This further yields,

$$\begin{aligned}
R \left(\hat{\Theta}_y^* \middle| y_{[N]} \right) &\geq \frac{|y|^2}{\sum_{n=1}^N \sum_{s,t \in y_n} \psi_{s,t} + |y|/\tau^2} \\
&= \frac{|y|^2}{\sum_{s,t \in y} N_{s \cup t} \psi_{s,t} + |y|/\tau^2}.
\end{aligned}$$

As a sanity check, when the sample size is zero ($N_{s \cup t} = 0, \forall s, t \in y$), we have $R(\hat{\Theta}_y^* | y_{[N]}) \geq |y|\tau^2$. This matches the intuition because if there is no historical data at all, the best estimator should just be the prior mean $|y|\mu$. It leads to the integrated risk $|y|\tau^2$ which contains no variance but only bias. \square

C Lemmas

Lemma 3. Let

$$S(q_s, N) = \sum_{N_s=1}^N \frac{1}{N_s} \binom{N}{N_s} q_s^{N_s} (1 - q_s)^{N - N_s}.$$

Then

$$\begin{aligned} \frac{1 - (1 - q_s)^{N+1}}{(N+1)q_s} - (1 - q_s)^N &< S(q_s, N) \\ 2 \left(\frac{1 - (1 - q_s)^{N+1}}{(N+1)q_s} - (1 - q_s)^N \right) &> S(q_s, N) \end{aligned}$$

Proof. For $N_s \sim \text{Binomial}(N, q_s)$, we use [Chao and Strawderman, 1972, Eqn. 3.4] to obtain

$$\begin{aligned} \mathbb{E}[(N_s + 1)^{-1}] &= (1 - q_s)^N + \sum_{N_s=1}^N \frac{1}{N_s + 1} \binom{N}{N_s} q_s^{N_s} (1 - q_s)^{N - N_s} \\ &= \frac{1 - (1 - q_s)^{N+1}}{(N+1)q_s}, \end{aligned}$$

and since

$$\begin{aligned} \sum_{N_s=1}^N \frac{1}{N_s + 1} \binom{N}{N_s} q_s^{N_s} (1 - q_s)^{N - N_s} &< S(q_s, N), \\ 2 \sum_{N_s=1}^N \frac{1}{N_s + 1} \binom{N}{N_s} q_s^{N_s} (1 - q_s)^{N - N_s} &> S(q_s, N), \end{aligned}$$

the result follows. □

Lemma 4. For an even number p , consider the following function of $j \in \mathbb{Z}_{>0}$,

$$g(j) = \prod_{i=1}^j \frac{i + \alpha}{i + 1} \cdot \frac{p - i}{p - i - 1 + \alpha},$$

we have,

$$g(j) \simeq \left(\frac{1}{j + 1} \right)^{1 - \alpha},$$

for all $0 \leq \alpha \leq 1$ and $j \leq p/2 - 1$.

Proof. We first show that $(i + \alpha)/(i + 1) \geq (i/(i + 1))^{1 - \alpha}$, $\forall i \in \mathbb{Z}_{>0}$. Notice that at the two end points $\alpha = 0$ and $\alpha = 1$, we have $(i + \alpha)/(i + 1) = (i/(i + 1))^{1 - \alpha}$, $\forall i \in \mathbb{Z}_{>0}$. Given that

$$\frac{d^2}{d\alpha^2} \left(\frac{i + \alpha}{i + 1} - \left(\frac{i}{i + 1} \right)^{1 - \alpha} \right) = - \left(\frac{i}{i + 1} \right)^{1 - \alpha} \log^2 \left(\frac{i}{i + 1} \right) < 0,$$

we know that $(i + \alpha)/(i + 1) - (i/(i + 1))^{1-\alpha}$ is concave in α for any $i \in \mathbb{Z}_{>0}$. This gives $(i + \alpha)/(i + 1) - (i/(i + 1))^{1-\alpha} \geq 0, \forall i \in \mathbb{Z}_{>0}, \forall \alpha \in [0, 1]$. We then have

$$g(j) \geq \prod_{i=1}^j \frac{i + \alpha}{i + 1} \geq \prod_{i=1}^j \left(\frac{i}{i + 1} \right)^{1-\alpha} = \left(\frac{1}{j + 1} \right)^{1-\alpha}.$$

We now prove the other direction. For some $i \in \mathbb{Z}_{>0}$, we first look at the function

$$f(\alpha) = \frac{(i + \alpha)/(i + 1)}{(i/(i + 1))^{1-\alpha}}, \quad \forall \alpha \in [0, 1].$$

The first-order condition of $f(\alpha)$ is

$$\frac{d}{d\alpha} f(\alpha) = \frac{\left(\frac{i}{i+1} \right)^\alpha \left((i + \alpha) \log\left(\frac{i}{i+1} \right) + 1 \right)}{i} = 0,$$

which yields

$$\alpha^*(i) = \frac{1}{\log((i + 1)/i)} - i \in [0, 1], \quad \forall i \in \mathbb{Z}_{>0}. \quad (30)$$

The second derivative of $f(\alpha)$ is

$$\begin{aligned} \frac{d^2}{d\alpha^2} f(\alpha) &= \frac{(i + \alpha) \left(\frac{i}{i+1} \right)^{\alpha-1} \log^2 \left(\frac{i}{i+1} \right)}{i + 1} + \frac{2 \left(\frac{i}{i+1} \right)^{\alpha-1} \log \left(\frac{i}{i+1} \right)}{i + 1} \\ &= \frac{1}{i + 1} \left(\frac{i}{i + 1} \right)^{\alpha-1} \log \left(\frac{i}{i + 1} \right) \left((i + \alpha) \log \left(\frac{i}{i + 1} \right) + 2 \right) < 0, \quad \forall \alpha \in [0, 1], \forall i \in \mathbb{Z}_{>0}. \end{aligned}$$

This suggests that $\alpha^*(i)$ in (30) is the solution that maximizes $f(\alpha)$ for a given $i \in \mathbb{Z}_{>0}$. Moreover, $\alpha^*(i)$ is increasing in i and $\lim_{i \rightarrow +\infty} \alpha^*(i) = 1/2$, which gives $\alpha^*(i) \in [0, 1/2], \forall i \in \mathbb{Z}_{>0}$.

We now show that

$$\prod_{i=1}^j \frac{i + \alpha}{i + 1} \lesssim \prod_{i=1}^j \left(\frac{i}{i + 1} \right)^{1-\alpha} = \left(\frac{1}{j + 1} \right)^{1-\alpha}, \quad \forall \alpha \in [0, 1], \forall j \in \mathbb{Z}_{>0}.$$

To see that,

$$\frac{\prod_{i=1}^j \frac{i + \alpha}{i + 1}}{\prod_{i=1}^j \left(\frac{i}{i + 1} \right)^{1-\alpha}} \leq \frac{\prod_{i=1}^j \frac{i + \alpha^*(i)}{i + 1}}{\prod_{i=1}^j \left(\frac{i}{i + 1} \right)^{1-\alpha^*(i)}} \leq \frac{\prod_{i=1}^j \frac{i + 1/2}{i + 1}}{\prod_{i=1}^j \left(\frac{i}{i + 1} \right)^{1-\alpha^*(i)}}. \quad (31)$$

For $\prod_{i=1}^j \left(\frac{i}{i + 1} \right)^{1-\alpha^*(i)}$,

$$\begin{aligned} &\prod_{i=1}^j \left(\frac{i}{i + 1} \right)^{1-\alpha^*(i)} \\ &= \prod_{i=1}^j \left(\frac{i}{i + 1} \right)^{1+i-\frac{1}{\log((i+1)/i)}} \end{aligned}$$

$$\begin{aligned}
&= \prod_{i=1}^j \left(\frac{i}{i+1}\right)^{1/2} \left(\frac{i}{i+1}\right)^{1/2+i} \left(\frac{i}{i+1}\right)^{-\frac{1}{\log((i+1)/i)}} \\
&= \prod_{i=1}^j \left(\frac{i}{i+1}\right)^{1/2} \left(\frac{i}{i+1}\right)^{1/2+i} e \\
&\geq \prod_{i=1}^j \left(\frac{i}{i+1}\right)^{1/2} \left(e^{-1} - \frac{1}{12e \cdot i^2}\right) e \quad (\text{using Taylor series of } (i/(i+1))^{1/2+i} \text{ at } +\infty) \\
&= \left(\frac{1}{j+1}\right)^{1/2} \prod_{i=1}^j \left(1 - \frac{1}{12 \cdot i^2}\right) \\
&\gtrsim \left(\frac{1}{j+1}\right)^{1/2}.
\end{aligned}$$

The last inequality uses a result in analysis that for a series of $0 < p_i < 1$, $i \in \mathbb{Z}_{>0}$, a sufficient and necessary condition for $\prod_{i=1}^{+\infty} (1 - p_i) > 0$ is $\sum_{i=1}^{+\infty} p_i < +\infty$. This leads to $\prod_{i=1}^j (1 - 1/(12 \cdot i^2)) \geq \prod_{i=1}^{+\infty} (1 - 1/(12 \cdot i^2)) > 0$.

On the other hand,

$$\prod_{i=1}^j \frac{i+1/2}{i+1} = \prod_{i=1}^j \frac{2i+1}{2i+2}.$$

We know that

$$\prod_{i=1}^j \frac{2i+1}{2i+2} \cdot \prod_{i=1}^j \frac{2i}{2i+1} = \frac{1}{j+1}, \quad \frac{\prod_{i=1}^j \frac{2i+1}{2i+2}}{\prod_{i=1}^j \frac{2i}{2i+1}} \leq (3/4)/(2/3) = 9/8,$$

which suggests that

$$\prod_{i=1}^j \frac{i+1/2}{i+1} \lesssim \left(\frac{1}{j+1}\right)^{1/2}.$$

Using (31),

$$\frac{\prod_{i=1}^j \frac{i+\alpha}{i+1}}{\prod_{i=1}^j \left(\frac{i}{i+1}\right)^{1-\alpha}} \lesssim \frac{\left(\frac{1}{j+1}\right)^{1/2}}{\left(\frac{1}{j+1}\right)^{1/2}} = 1 \quad \Rightarrow \quad \prod_{i=1}^j \frac{i+\alpha}{i+1} \lesssim \prod_{i=1}^j \left(\frac{i}{i+1}\right)^{1-\alpha} = \left(\frac{1}{j+1}\right)^{1-\alpha}.$$

This further gives,

$$\begin{aligned}
g(j) &= \prod_{i=1}^j \frac{i+\alpha}{i+1} \cdot \frac{p-i}{p-i-1+\alpha} \\
&\leq \prod_{i=1}^j \frac{i+\alpha}{i+1} \cdot \frac{p-i}{p-i-1} \\
&= \left(\prod_{i=1}^j \frac{i+\alpha}{i+1}\right) \cdot \frac{p-1}{p-j-1}
\end{aligned}$$

$$\begin{aligned}
&\leq \left(\prod_{i=1}^j \frac{i+\alpha}{i+1} \right) \cdot \frac{p-1}{p-p/2+1-1} \\
&= \left(\prod_{i=1}^j \frac{i+\alpha}{i+1} \right) \cdot \frac{p-1}{p/2} \\
&\leq 2 \left(\prod_{i=1}^j \frac{i+\alpha}{i+1} \right) \\
&\lesssim \left(\frac{1}{j+1} \right)^{1-\alpha}.
\end{aligned}$$

This completes the proof. \square

Lemma 5. Under the route distribution μ_p in Section 3.1,

$$\sum_{s \in \mathcal{S}_p} q_s^2 = \sum_{s \in \mathcal{S}_p} \mathbb{P}^2 [s \in Y_p] \simeq \begin{cases} 1, & \frac{1}{2} < \alpha \leq 1, \\ p^{1-2\alpha}, & 0 < \alpha \leq \frac{1}{2}. \end{cases}$$

Proof. Without loss of generality, we focus on the case that p is even. From the proof of Proposition 4, we know that for a segment $s = (i, j) \rightarrow (i+1, j)$,

$$\begin{aligned}
&\mathbb{P} [s \in Y_p] \\
&\simeq \sum_{i_1=0}^i \sum_{i_2=i+1}^p \mathbb{P}(x_1 = (i_1, j)) \mathbb{P}(x_2 = (i_2, \cdot)) \\
&= \mathbb{P}(x_1 = (\cdot, j)) \cdot \left(\sum_{i_1=0}^i \sum_{i_2=i+1}^p \mathbb{P}(x_1 = (i_1, \cdot)) \mathbb{P}(x_2 = (i_2, \cdot)) \right) \\
&= \mathbb{P}(x_1 = (\cdot, j)) \cdot \left(\sum_{i_1=0}^i \mathbb{P}(x_1 = (i_1, \cdot)) \right) \left(\sum_{i_2=i+1}^p \mathbb{P}(x_2 = (i_2, \cdot)) \right).
\end{aligned}$$

Moreover, because μ_p is symmetric,

$$\begin{aligned}
&\sum_{s \in \mathcal{S}_p} \mathbb{P}^2 [s \in Y_p] \\
&\simeq \sum_{i \in \{0, \dots, p/2-1\}} \sum_{j \in \{0, \dots, p/2\}} \mathbb{P}^2 [s = (i, j) \rightarrow (i+1, j) \in Y_p] \\
&\simeq \sum_{i \in \{0, \dots, p/2-1\}} \sum_{j \in \{0, \dots, p/2\}} \mathbb{P}^2(x_1 = (\cdot, j)) \cdot \left(\sum_{i_1=0}^i \mathbb{P}(x_1 = (i_1, \cdot)) \right)^2 \left(\sum_{i_2=i+1}^p \mathbb{P}(x_2 = (i_2, \cdot)) \right)^2 \\
&= \left(\sum_{j \in \{0, \dots, p/2\}} \mathbb{P}^2(x_1 = (\cdot, j)) \right) \cdot \left(\sum_{i \in \{0, \dots, p/2-1\}} \left(\sum_{i_1=0}^i \mathbb{P}(x_1 = (i_1, \cdot)) \right)^2 \left(\sum_{i_2=i+1}^p \mathbb{P}(x_2 = (i_2, \cdot)) \right)^2 \right) \\
&\simeq \left(\sum_{j \in \{0, \dots, p/2\}} \mathbb{P}^2(x_1 = (\cdot, j)) \right) \cdot \left(\sum_{i \in \{0, \dots, p/2-1\}} \left(\sum_{i_1=0}^i \mathbb{P}(x_1 = (i_1, \cdot)) \right)^2 \right). \tag{32}
\end{aligned}$$

- When $1/2 < \alpha \leq 1$, from the proof of Proposition 4,

$$\sum_{j \in \{0, \dots, p/2\}} \mathbb{P}^2(x_1 = (\cdot, j)) \simeq p^{-1},$$

and

$$\sum_{i_1=0}^i \mathbb{P}(x_1 = (i_1, \cdot)) \simeq p^{-\alpha} \left(\sum_{i_1=0}^i \left(\frac{1}{i_1+1} \right)^{1-\alpha} \right) \simeq p^{-\alpha} (i+1)^\alpha.$$

This yields,

$$(32) \simeq p^{-1} p^{-2\alpha} \sum_{i \in \{1, \dots, p/2\}} i^{2\alpha} \simeq p^{-1} p^{-2\alpha} p^{2\alpha+1} \simeq 1.$$

- When $0 < \alpha \leq 1/2$, from the proof of Proposition 4,

$$\sum_{j \in \{0, \dots, p/2\}} \mathbb{P}^2(x_1 = (\cdot, j)) \simeq p^{-2\alpha},$$

and

$$\sum_{i_1=0}^i \mathbb{P}(x_1 = (i_1, \cdot)) \simeq p^{-\alpha} \left(\sum_{i_1=0}^i \left(\frac{1}{i_1+1} \right)^{1-\alpha} \right) \simeq p^{-\alpha} (i+1)^\alpha.$$

This yields,

$$(32) \simeq p^{-2\alpha} p^{-2\alpha} \sum_{i \in \{1, \dots, p/2\}} i^{2\alpha} \simeq p^{-2\alpha} p^{-2\alpha} p^{2\alpha+1} \simeq p^{1-2\alpha}.$$

This completes the proof. □

Lemma 6. For any $i, j \in \{0, \dots, p\}$,

$$\sum_{i_1=0}^i \sum_{i_2=i+1}^p \sum_{j_2=0}^p \frac{1}{(|i_1 - i_2| + |j - j_2|)^2} \leq 2 \left(\sum_{n=1}^{i+1} \frac{1 + \dots + n}{n^2} + \sum_{n=i+2}^{2p} \frac{1 + \dots + (i+1) + (i+1)(n-i-1)}{n^2} \right).$$

Proof. We have,

$$\begin{aligned} & \sum_{i_1=0}^i \sum_{i_2=i+1}^p \sum_{j_2=0}^p \frac{1}{(|i_1 - i_2| + |j - j_2|)^2} \\ &= \sum_{j_2=0}^p \sum_{i_2=i+1}^p \sum_{i_1=i}^0 \frac{1}{(|i_1 - i_2| + |j - j_2|)^2} \\ &\leq 2 \left(\sum_{j_2=0}^p \sum_{i_2=i+1}^p \sum_{i_1=i}^0 \frac{1}{(|i_1 - i_2| + j_2)^2} \right) \quad (\text{by symmetry and } j \in \{0, \dots, p\}) \end{aligned}$$

$$\begin{aligned}
&= 2 \left(\sum_{j_2=0}^p \sum_{i_2=1}^{p-i} \sum_{i_1=0}^i \frac{1}{(i_1 + i_2 + j_2)^2} \right) \\
&\leq 2 \left(\sum_{n=1}^{i+1} \frac{1 + \dots + n}{n^2} + \sum_{n=i+2}^{2p} \frac{1 + \dots + (i+1) + (i+1)(n-i-1)}{n^2} \right).
\end{aligned}$$

This completes the proof. \square

Lemma 7. Under the route distribution μ_p in Section 3.1, for any segment $s \in \mathcal{S}_p$,

$$\mathbb{E} \left[\frac{1}{(\sum_{s' \in \mathcal{S}_p} I_{s'})^2} \middle| I_s = 1 \right] = \mathcal{O}(\log(p)p^{-2}), \quad \mathbb{E} \left[\frac{1}{(\sum_{s' \in \mathcal{S}_p} I_{s'})^2} \middle| I_s = 0 \right] = \mathcal{O}(\log(p)p^{-2}).$$

Proof. For any segment s , $\mathbb{E}[1/(\sum_{s' \in \mathcal{S}_p} I_{s'})^2 | I_s = 1]$ and $\mathbb{E}[1/(\sum_{s' \in \mathcal{S}_p} I_{s'})^2 | I_s = 0]$ are increasing in $\alpha \in (0, 1]$, this is simply because as α increases, route origins and destinations are more concentrated in the center of the grid. We can thus focus on the case where $\alpha = 1$ to get upper bounds. Let \mathcal{X}_s be the set of origins and destinations such that $\mathbb{P}[x_1 = (i_1, j_1), x_2 = (i_2, j_2) | I_s = 1] > 0$ for any origin-destination pair $((i_1, j_1), (i_2, j_2)) \in \mathcal{X}_s$. Similarly, let \mathcal{X}'_s be the set of origins and destinations such that $\mathbb{P}[x_1 = (i_1, j_1), x_2 = (i_2, j_2) | I_s = 0] > 0$ for any origin-destination pair $((i_1, j_1), (i_2, j_2)) \in \mathcal{X}'_s$.

For any segment $s \in \mathcal{S}_p$, when $\alpha = 1$, i.e., route origins and destinations are uniformly distributed over the grid,

$$\begin{aligned}
&\mathbb{E} \left[\frac{1}{(\sum_{s' \in \mathcal{S}_p} I_{s'})^2} \middle| I_s = 1 \right] \\
&= \sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}_s} \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \cdot \mathbb{P}[x_1 = (i_1, j_1), x_2 = (i_2, j_2) | I_s = 1] \\
&= \sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}_s} \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \cdot \frac{\mathbb{P}[I_s = 1 | x_1 = (i_1, j_1), x_2 = (i_2, j_2)] \mathbb{P}[x_1 = (i_1, j_1), x_2 = (i_2, j_2)]}{\mathbb{P}[I_s = 1]} \\
&\simeq \sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}_s} \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \cdot \frac{\mathbb{P}[x_1 = (i_1, j_1), x_2 = (i_2, j_2)]}{\sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}_s} \mathbb{P}[x_1 = (i_1, j_1), x_2 = (i_2, j_2)]} \tag{33}
\end{aligned}$$

$$= \sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}_s} \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \cdot \frac{1}{|\mathcal{X}_s|}. \tag{34}$$

Equation (33) holds because $\mathbb{P}[I_s = 1 | x_1 = (i_1, j_1), x_2 = (i_2, j_2)] \in \{0.5, 1\}$, $\forall ((i_1, j_1), (i_2, j_2)) \in \mathcal{X}_s$. Moreover, equation (34) holds because of the uniformity of the distribution of origins and destinations.

Without loss of generality, consider any segment with horizontal movement $s = (i, j) \rightarrow (i + 1, j) \in \mathcal{S}_p$ with $i < p/2$,

$$\begin{aligned}
&\mathbb{E} \left[\frac{1}{(\sum_{s' \in \mathcal{S}_p} I_{s'})^2} \middle| I_s = 1 \right] \\
&\simeq \sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}_s} \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \cdot \frac{1}{|\mathcal{X}_s|}
\end{aligned}$$

$$\begin{aligned}
&= \sum_{i_1 \in \{0, \dots, i\}} \sum_{i_2 \in \{i+1, \dots, p\}} \sum_{j_2 \in \{0, \dots, p\}} \frac{1}{(|i_1 - i_2| + |j - j_2|)^2} \cdot \frac{1}{|\mathcal{X}_s|} \\
&\quad + \sum_{i_1 \in \{0, \dots, i\}} \sum_{j_1 \in \{0, \dots, p\}} \sum_{i_2 \in \{i+1, \dots, p\}} \frac{1}{(|i_1 - i_2| + |j_1 - j|)^2} \cdot \frac{1}{|\mathcal{X}_s|} \\
&= \frac{2}{|\mathcal{X}_s|} \left(\sum_{i_1 \in \{0, \dots, i\}} \sum_{i_2 \in \{i+1, \dots, p\}} \sum_{j_2 \in \{0, \dots, p\}} \frac{1}{(|i_1 - i_2| + |j - j_2|)^2} \right) \\
&\leq \frac{4}{|\mathcal{X}_s|} \left(\sum_{n=1}^{i+1} \frac{1 + \dots + n}{n^2} + \sum_{n=i+2}^{2p} \frac{1 + \dots + (i+1) + (i+1)(n-i-1)}{n^2} \right) \quad (\text{by Lemma 6 in Appendix C}) \\
&= \frac{4}{|\mathcal{X}_s|} \left(\sum_{n=1}^{i+1} \frac{n(n+1)/2}{n^2} + \sum_{n=i+2}^{2p} \frac{(i+2)(i+1)/2 + (i+1)(n-i-1)}{n^2} \right) \\
&\leq \frac{4}{|\mathcal{X}_s|} \left((i+1) + (i+1) \sum_{n=i+2}^{2p} \frac{n-i/2}{n^2} \right) \\
&= \frac{4}{|\mathcal{X}_s|} (i+1) \mathcal{O}(\log(p)) \\
&\simeq \frac{4}{(i+1)p^2} (i+1) \mathcal{O}(\log(p)) \\
&= \mathcal{O}(\log(p)p^{-2}).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{E} \left[\frac{1}{(\sum_{s' \in \mathcal{S}_p} I_{s'})^2} \middle| I_s = 0 \right] \\
&= \sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}'_s} \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \cdot \mathbb{P}[x_1 = (i_1, j_1), x_2 = (i_2, j_2) \mid I_s = 0] \\
&= \sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}'_s} \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \cdot \frac{\mathbb{P}[I_s = 0 \mid x_1 = (i_1, j_1), x_2 = (i_2, j_2)] \mathbb{P}[x_1 = (i_1, j_1), x_2 = (i_2, j_2)]}{\mathbb{P}[I_s = 0]} \\
&\simeq \sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}'_s} \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \cdot \frac{\mathbb{P}[x_1 = (i_1, j_1), x_2 = (i_2, j_2)]}{\sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}'_s} \mathbb{P}[x_1 = (i_1, j_1), x_2 = (i_2, j_2)]} \quad (35)
\end{aligned}$$

$$= \sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}'_s} \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \cdot \frac{1}{|\mathcal{X}'_s|}. \quad (36)$$

Equation (35) holds because $\mathbb{P}[I_s = 0 \mid x_1 = (i_1, j_1), x_2 = (i_2, j_2)] \in \{0.5, 1\}$, $\forall ((i_1, j_1), (i_2, j_2)) \in \mathcal{X}'_s$, and equation (36) holds because of the uniformity of the distribution of origins and destinations.

Without loss of generality, consider any segment with horizontal movement $s = (i, j) \rightarrow (i+1, j) \in \mathcal{S}_p$ with $i < p/2$,

$$\mathbb{E} \left[\frac{1}{(\sum_{s' \in \mathcal{S}_p} I_{s'})^2} \middle| I_s = 0 \right]$$

$$\begin{aligned}
&\simeq \sum_{((i_1, j_1), (i_2, j_2)) \in \mathcal{X}'_s} \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \cdot \frac{1}{|\mathcal{X}'_s|} \\
&\simeq \left(\sum_{i_1=0}^p \sum_{i_2=0}^p \sum_{j_1=0}^p \sum_{j_2=0}^p \frac{1}{(|i_1 - i_2| + |j_1 - j_2|)^2} \right) \cdot \frac{1}{p^4} \\
&= \mathcal{O} \left((p^2 \log(p)) \cdot \frac{1}{p^4} \right) \\
&= \mathcal{O}(p^{-2} \log(p)).
\end{aligned}$$

The second equality holds because the only origin-destination pairs that are *excluded* in \mathcal{X}'_s are those with $j_1 = j_2 = j$ and $i_1 \in \{0, \dots, i\}, i_2 \in \{i + 1, \dots, p\}$. The cardinality of these origin-destination pairs is of a much smaller order (p^2) compared to the cardinality of all possible origin-destination pairs (of the order of p^4). This completes the proof. \square

D Additional Numerical Experiments

Figure 7 and Figure 8 report additional numerical experiments based on the setup in Section 3.2 under $\alpha = 1.0$ with two different specifications of the covariance matrices $e^{-3\mathcal{L}} + I$ and $3e^{-\mathcal{L}} + I$. The results are qualitatively similar. Simple segment-based method tends to perform a bit worse when correlation is strong and sample size is small, but it quickly regains competitiveness as the sample size increases. We further test two additional covariance structures which do not satisfy the second part of Assumption 4, under $\alpha = 1.0$. In both cases, the covariance matrices of the segment travel times Σ_p for the grid network with size p is constructed as $\Sigma_p = (1/|\mathcal{S}_p|^2) K_p^\top K_p$ where K_p is an $|\mathcal{S}_p| \times |\mathcal{S}_p|$ random matrix. The two cases differ in the way K_p is generated.

D.1 Entries in K_p are drawn from $\mathcal{U}_{[-1,1]}$

In the first case, K_p is a random matrix whose elements are drawn from a uniform distribution between $[-1, 1]$. It can be checked that both the covariance matrix $\Sigma_p = (1/|\mathcal{S}_p|^2) K_p^\top K_p$ and the precision matrix $\Psi_p = \Sigma_p^{-1}$ violate the second part of Assumption 4. The variance of the means of segment travel times τ^2 is set to be 0.5 which is similar to the variance of the segment travel times σ_s^2 . The rest of the experimental setups are the same as those in Section 3.2. Figure 9 shows similar trends for the integrated risks of the simple segment-based estimators and the optimal route-based estimators, as those in Figure 4. This is not too much out of expectation as the proof of Theorem 5 and Theorem 3 do not require spatial decay of precision matrix Ψ_p . Moreover, the entries in the covariance matrix are mostly dominated by the diagonal and the off-diagonal entries sum up to zero in expectation. This roughly gives $\sum_{s,t \in \mathcal{S}_p} \sigma_{s,t} = \mathcal{O}(|\mathcal{S}_p|) = \mathcal{O}(p^2)$.

On the other hand, the information-theoretic lower bounds seem to be of lower order compared to the integrated risks of the simple segment-based estimator. This is expected as the proof of Theorem 4 critically uses the spatial decay of the precision matrix Ψ_p . What is surprising is that the simple segment-based estimator is still highly competitive compared to the optimal estimator — its risk is very close to the optimal risk.

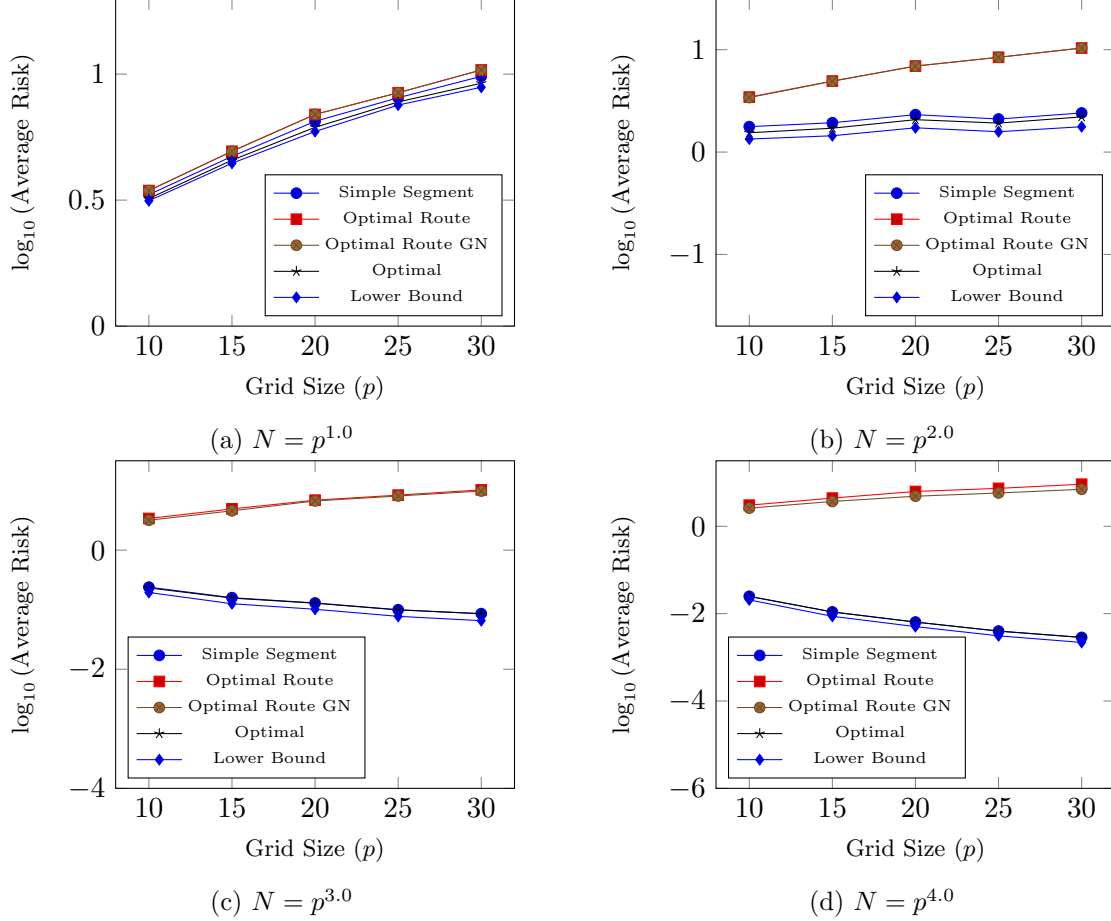


Figure 7: Average integrated risks of different estimators ($\alpha = 1.0$, covariance is $e^{-3\mathcal{L}} + I$).

D.2 Entries in K_p are drawn from $\mathcal{U}_{[0,1]}$

In the second case, K_p is a random matrix whose elements are drawn from a uniform distribution between $[0, 1]$. It can be checked that both the covariance matrix $\Sigma_p = (1/|\mathcal{S}_p|^2)K_p^\top K_p$ and the precision matrix $\Psi_p = \Sigma_p^{-1}$ violate the second part of Assumption 4. In particular, we have $\sum_{s,t \in \mathcal{S}_p} |\sigma_{s,t}| = \sum_{s,t \in \mathcal{S}_p} \sigma_{s,t} = \mathcal{O}(|\mathcal{S}_p|^2)$. The variance of the means of segment travel times τ^2 is set to be 0.5 which is similar to the variance of the segment travel times σ_s^2 . Figure 9 now shows quite different trends for the integrated risks of the simple segment-based estimators and the optimal route-based estimators, compared to those in Figure 4. Following the same proof of Theorem 5, one can show that $R(\hat{\Theta}_{Y_p}^{(\text{seg})}) = \mathcal{O}(|\mathcal{S}_p|^2/N) = \mathcal{O}(p^4/N)$ which is reflected by Figures 10. The simple segment-based estimator with $\phi_s(N_s) = N_s/(N_s + 1)$, $\forall s \in y$ weighs too much on the training data, and since segment travel times in this setting have much higher variance, the optimal segment-based estimator places more weights on the prior mean when the sample size is small. This results in poor performance of the simple segment-based estimators (outperformed by the optimal route-based estimator) when sample size is small. The performance of the simple segment-based estimator improves significantly as the sample size gets larger, and eventually dominates the optimal route-based estimator.

Similarly to Figure 9, the information-theoretic lower bounds are of lower order compared to the risks of the simple segment-based estimator. The difference is that the gap between the simple

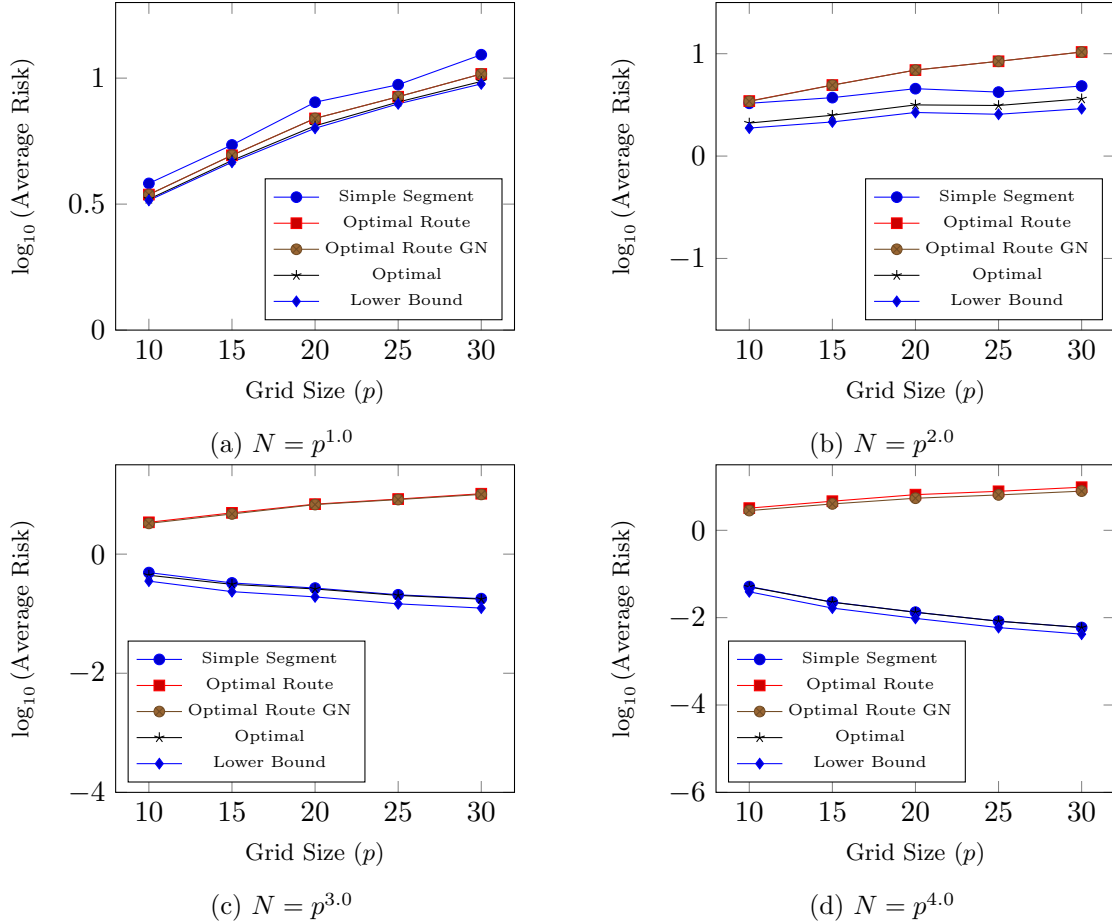


Figure 8: Average integrated risks of different estimators ($\alpha = 1.0$, covariance is $3e^{-\mathcal{L}} + I$).

segment-based estimator and the optimal estimator gets larger. These observations along with those in Section D.2 suggest that controlling the covariance matrix of the segment travel times is likely much more important than controlling the precision matrix in maintaining the optimality of the simple segment-based estimator.

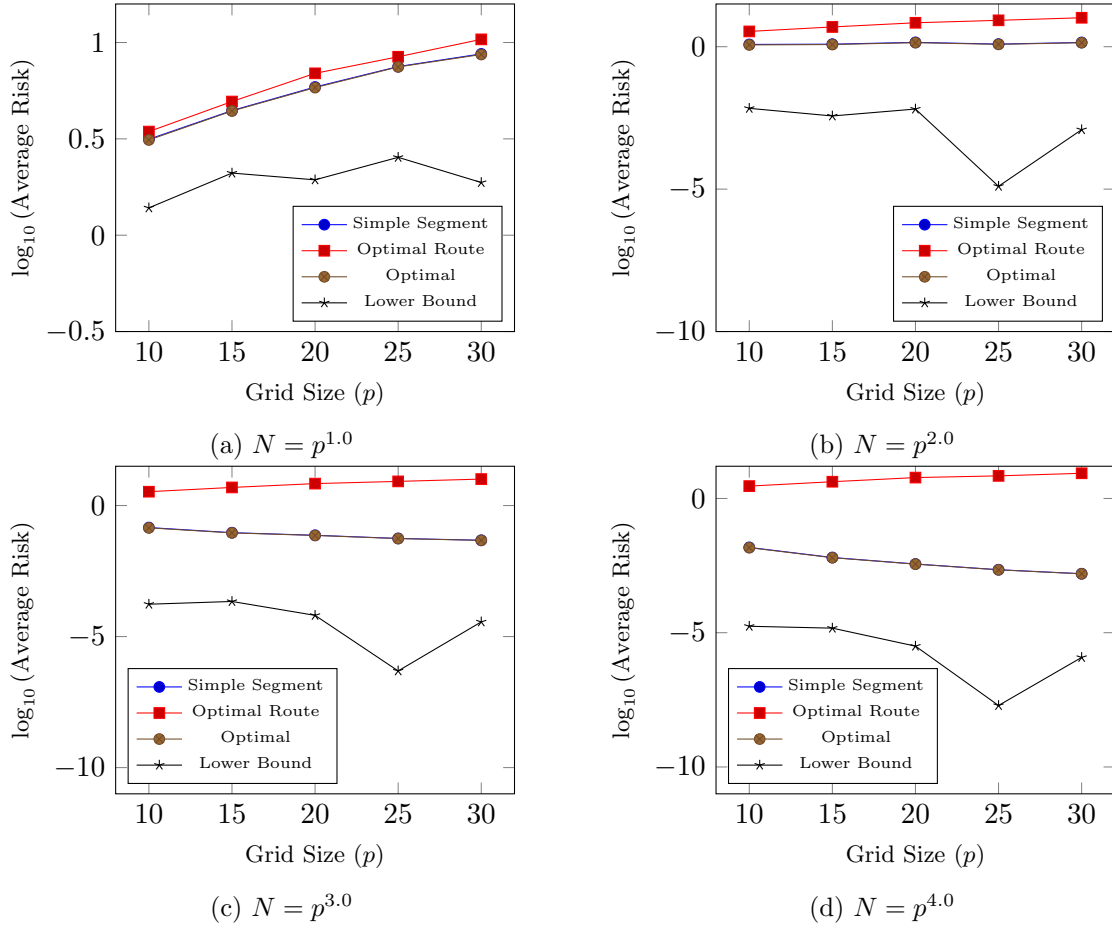


Figure 9: Average integrated risks of different estimators ($\alpha = 1.0$). Covariance of segment travel times is $\Sigma_p = K_p^\top K_p / |\mathcal{S}_p|^2$ where K_p is a random matrix whose entries are drawn from $\mathcal{U}_{[-1,1]}$.

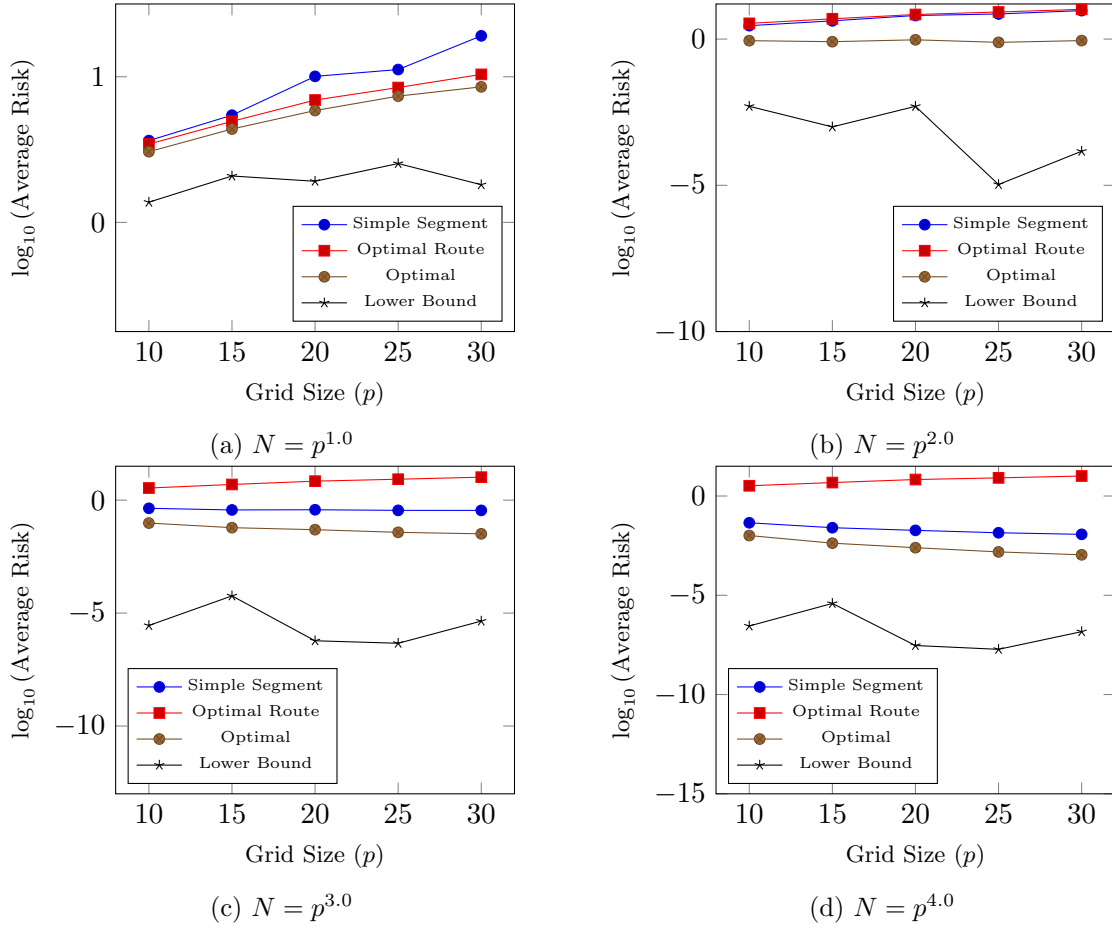


Figure 10: Average integrated risks of different estimators ($\alpha = 1.0$). Covariance of segment travel times is $\Sigma_p = K_p^\top K_p / |\mathcal{S}_p|^2$ where K_p is a random matrix whose entries are drawn from $\mathcal{U}_{[0,1]}$.