University of Warsaw
Faculty of Mathematics, Informatics and Mechanics

Jędrzej Jabłoński

Structured population models for predator-prey interactions. The case of *Daphnia* and size-selective planktivorous fish.

PhD Dissertation

Supervised by
prof. dr hab. Dariusz Wrzosek

June 2014

Supervisor's statement

This dissertation is ready to be reviewed.

Date Supervisor's signature

Authors' statement

Hereby I declare that the present thesis was prepared by myself and none of its contents was obtained by means that are against the law. The thesis has never before been a subject of any procedure of obtaining academic degree. Moreover, I declare that the present version of the thesis is identical with the attached electronic version.

Date Authors' signatures

Contents

Abstract

In this thesis a model of the dynamics of size-structured population subject to selective predation is built and analyzed. The study is motivated by biological phenomena concerning limnology and oceanography, and in particular diversity of first consumers in aquatic ecosystems. An individual-based model of size-selective visual predator-harvester based on the concept of optimal foraging is proposed. Farther, a simplification of the model, described in terms of operators on the space of measures, is derived based on Holling II-type functional response to eliminate inherent difficulties of individual-based approach. The results are compared against experimental evidence. Considerations involving populations dynamics, namely growth, birth and mortality, are examined in the framework of measure-valued solutions to transport equation and various distances arising from optimal transportation theory. To this end, efficient algorithms for solving transportation problem on a real line are found and finally, numerical schemes based on particle methods for structured population models are improved. Moreover, approximation theory for Radon measures is developed.

Acknowledgments

I would like to thank my supervisor, prof. Dariusz Wrzosek from the Institute of Applied Mathematics and Mechanics at University of Warsaw, for collaboration, inspiring discussions, and many valuable comments. I am also grateful for the opportunity of working with prof. Anna Marciniak-Czochra from Heidelberg University.

In addition, I have been privileged to collaborate with prof. Piotr Gwiazda, dr. Agnieszka Ulikowska from the Institute of Applied Mathematics and Mechanics and prof. Z. Maciej Gliwicz, and mgr. Piotr Maszczyk from the Department of Hydrobiology at University of Warsaw. Long discussions with you have significantly improved my work and inspired many new research directions.

I was supported by the International Ph.D. Projects Programme of Foundation for Polish Science operated within the Innovative Economy Operational Programme 2007-2013 funded by European Regional Development Fund (Ph.D. Programme: Mathematical Methods in Natural Sciences).

Notation

In this thesis the following notation is used:

- $\mathbb{R}^{\geq 0}$ is the set of non-negative real numbers,
- $\mathbb{R}^{+}$ is the set of positive real numbers,
- $Lip(f)$ is the Lipschitz constant of function f,
- μ^{+} is the non-negative measure arising from Jordan decomposition of μ,
- $D_{\nu}\mu$ is the Radon-Nikodym derivative of measure μ with respect to ν,
- $\mathcal{L}$ is the Lebesgue measure,
- $\mathcal{O}, \Theta, \Omega$ is the standard Landau notation for limiting behavior,
- $\mathbb{1}_{E}$ is an indicator function of set E,
- $\mu|_{E}$ is the restriction of measure μ to the set E,
- $C_1, C_2, ...$ are absolute constants that may differ between occurrences.

For normed spaces X and Y we shall use following notation:

- $C(X;Y)$ is the space of continuous functions,
- $C_b(X;Y)$ is the space of bounded continuous functions,
- $C^{0,1}(X;Y)$ is the space of Lipschitz continuous functions,
- $C_0(X;Y)$ is the space of continuous function vanishing at infinity,
- $C_c(X;Y)$ is the space of compactly supported continuous function,
- $L^p(X;Y)$ is the usual Lebesgue space,
- $\mathfrak{B}(X)$ is the Borel σ-algebra on X,
- $\mathfrak{M}(X)$ is the space of finite, Radon measures,
- $\mathfrak{M}_d(X)$ is a subset of $\mathfrak{M}(X)$ consisting of discrete measures with finite number of atoms,
- $\mathfrak{M}_{d,N}(X)$ is a subset of $\mathfrak{M}_d(X)$ consisting of discrete measures with N atoms,
- $\langle \mu, f \rangle$ for measure $\mu \in \mathfrak{M}(X)$ and function $f \in C(X;\mathbb{R})$ is the value $\int_X f d\mu$,
- $B_X(x,r)$ is the set $\{y \in X : \|x-y\|_X \leq r\}$.

For simplicity notation $\mathfrak{M}[a,b]$ and $L^p[a,b]$ is often used instead of $\mathfrak{M}([a,b])$ and $L^p([a,b])$. Similarly, notation $C(X)$ is used instead of $C(X;\mathbb{R})$.

If $\gamma \in \mathfrak{M}(X \times X)$ then for a given set $A \subseteq X$ we define measure $\gamma(A,\cdot) \in \mathfrak{M}(X)$ by $\gamma(A,\cdot)(E) = \gamma(A \times E)$ for every measurable set $E \subseteq X$.

Introduction

The goal of this thesis is to build and analyze a model of size-structured population subject to selective predation. The study is motivated by biological phenomena concerning limnology and oceanography, and in particular diversity of first consumers in aquatic ecosystems. An individual-based model of size-selective visual predator-harvester based on the concept of optimal foraging is proposed [41]. It incorporates models of underlying physical processes and makes predictions based on the assumption that the forager maximizes its rate of energy intake [81, 71, 65, 59, 52, 8, 84]. Farther, a simplification of the model is derived to eliminate inherent difficulties of individual-based approach. A generalization of Holling II-type model [38] is proposed and the results are compared against experimental evidence collected by a team of hydrobiologists affiliated with the University of Warsaw [28, 29, 53, 30]. Considerations involving populations dynamics, namely growth, birth and mortality, are examined in the framework of measure-valued solutions [35, 36] to transport equation [2] and optimal transportation theory [76]. To this end, the theory of approximation on the space of finite Radon measures equipped with bounded Lipschitz distance is developed, efficient algorithms for solving transportation problem on a real line are found [40] and finally, numerical schemes based on particle methods for structured population models are improved.

The dissertation is divided into three almost independent parts treating theory of metrics on the space of measures, theory of measure-valued McKendrick-von Foerster equations and optimal foraging models. This order has been chosen for the convenience of a reader with mathematical background. The main results of the first chapter consist of an algorithm for computing bounded Lipschitz distance between two discrete measures supported on an N-element subset of $\mathbb{R}$. Computational complexity of this algorithm is proved to be $\mathcal{O}(N \log N)$. Moreover, a number of theorems characterizing optimal approximations of different classes of measures by discrete measures supported on an N-element set are proved [39]. In the second chapter well-established numerical schemes based on particle methods [15], such as split-up algorithm, original escalator box-car train and its modification are compared [34] and three improvements basing on the results of the previous chapter are described. Moreover, it is demonstrated that a certain generalization of Holling II-type model of foraging can be translated into the language of operators on spaces of measures, and since appropriate regularity conditions hold it can be used in McKendrick-von Foerster population dynamics equations. The last chapter describes three novel models of size-selective visual predator-harvester feeding on a prey population homogeneously distributed in space based on the concept of optimal foraging [41]. Optimization of the rate of net energy intake occurs at the level of forager's decisions, which include cruising speed [79, 78, 66], attack velocity and active selection of prey items [49, 26, 20]. The greatest advantage of models proposed in this chapter is that all parameters are physically measurable and no fitting to experimental data is required. Finally, the outcome of model simulations is compared against experimental data, collected by the hydrobiologists, and critically discussed. The thesis and the proposed models improves comprehension of many aspects of foraging in an aquatic environment.

Chapter 1

Metrics on the spaces of Radon measures

The space of finite Radon measures on X, $\mathfrak{M}(X)$, is naturally equipped with a norm induced by the total variation, which makes $\mathfrak{M}(X)$ a Banach space. However, the metric induced by this norm is so strong that it does not provide a reasonable measure of error for most applications. For instance, it is often desired that two Dirac masses with atoms close to each other in X are also close in some metric on $\mathfrak{M}(X)$. For this reason a different notion of distance has to be developed.

In many applications such as transportation problems [44, 76], crowd dynamics [54, 55], structured population dynamics [11, 35, 36, 73] or gradient flows [5, 82] it is natural to consider the output of mathematical modeling in terms of Radon measures, rather than densities. One reason is that very basic phenomena (e.g. growth of individuals in structured population models) may lead to singularities in density functions. What seems to be even more important is that mathematical tools used for the analysis of function-valued solutions (as opposed to measure-valued solutions) imply an inherently inappropriate sense of distance between solutions (see Example 10). The desired properties of such distance depend on the structure of the considered problem [36]. Recent years witnessed large developments in the kinetic theory methods applied to mathematical physics and more recently also to mathematical biology. Among important branches of the kinetic theory are optimal transportation problems and related to them Wasserstein metrics or Monge-Kantorovich metrics [5, 76]. These, however, are only applicable to processes with mass conservation. To cope with variable mass, several modifications have been proposed, including flat metric, centralized Wasserstein metric and normalized Wasserstein distance. For comparison of different metrics, their interpretation and examples we refer to Section 1.4.

Metrics based on the concept of optimal transportation have been used in different fields such as image recognition [24], alignment of surfaces [50], fluid dynamics [32], asymptotics of nonlinear diffusion equations [13], semi-supervised learning [69]. This chapter is mainly devoted to the flat metric, which is a natural choice for population models studied in chapter 2.

1.1. Preliminaries

Throughout this chapter we assume that X is a finite-dimensional Banach space. Some definitions and results can be generalized to locally compact metric spaces. It is, however, beyond the scope of this chapter.

Definition 1. Mapping $\mu : \mathfrak{B}(X) \to \mathbb{R} \cup \{-\infty, \infty\}$ is called a Radon measure if the following conditions hold

1. $\mu(\emptyset) = 0$
2. for any countable collection, $\{E_i\}_{i=1}^{\infty} \subset \mathfrak{B}(X)$, of pairwise disjoint sets
$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mu\left(E_i\right),$$
3. μ takes at most one of the values $-\infty$ and ∞.

Definition 2. Let μ be a Radon measure on X. By total variation of μ we mean
$$\|\mu\| = \mu^+(X) - \mu^-(X).$$

Existence and unique decomposition of arbitrary Radon measure μ into a difference of two non-negative measures μ^+ and μ^- follows from Jordan decomposition theorem. Measures with finite total variation are called finite. The set of all finite Radon measures on X are be denoted by $\mathfrak{M}(X)$.

Definition 3. We define the space of bounded Lipschitz functions as
$$C_b^{0,1}(X;Y) = C^{0,1}(X;Y) \cap C_b(X;Y)$$
equipped with the following norm
$$\|f\|_{C_b^{0,1}(X;Y)} = \max\left(\|f\|_{C(X;Y)}, \sup_{x,y \in X} \frac{\|f(x) - f(y)\|_Y}{\|x - y\|_X}\right).$$
The norm $\|\cdot\|_{C_b^{0,1}(X;Y)}$ is known as the Fortet-Mourier norm (see [23]).

Theorem 4. *(Riesz-Markov representation theorem) Let $\psi \in C_0(X)^*$ then there exists a unique $\mu \in \mathfrak{M}(X)$ such that for every $f \in C_0(X)$*
$$\psi(f) = \int_X f d\mu.$$

Theorem 5. *(Riesz-Markov-Kakutani representation theorem) Let $\psi \in C_c(X)^*$ then there exists a unique Radon measure, μ, on X such that for every $f \in C_c(X)$*
$$\psi(f) = \int_X f d\mu.$$

Definition 6. We define the following norms on some subspaces of $\mathfrak{M}(X)$:

$$\begin{aligned}
\|\mu\|_R &= \sup\left\{\int_X f d\mu : f \in B_{C(X)}(0,1)\right\},\\
\|\mu\|_W &= \sup\left\{\int_X f d\mu : f \in C^{0,1}(X),\, Lip(f) \le 1\right\},\\
\|\mu\|_F &= \sup\left\{\int_X f d\mu : f \in B_{C_b^{0,1}(X)}(0,1)\right\}.
\end{aligned}$$

Proposition 7. *Let $\Omega \subset X$ be a compact set and let $\mu \in \mathfrak{M}(X)$, then $\|\mu\| = \|\mu\|_R \ge \|\mu\|_F$ and $\|\mu\|_W \ge \|\mu\|_F$.*

Definition 8. For each of the defined norms we define corresponding metrics: Radon distance, 1-Wasserstein distance and flat distance

$$\begin{aligned}
\rho(\mu,\nu) &= \|\mu-\nu\|_R,\\
W(\mu,\nu) &= \|\mu-\nu\|_W,\\
\rho_F(\mu,\nu) &= \|\mu-\nu\|_F.
\end{aligned}$$

Proposition 9. *Let $\Omega \subset X$ be a compact set, and let $\mu,\nu \in \mathfrak{M}^+(X)$. Then, $W(\mu,\nu) < \infty$ if and only if $\|\mu\| = \|\nu\|$.*

Proof. Let $K = \sup_{x\in X}\|x\|_X$. Choose a sequence $\{f_n\} \subset C^{0,1}(X)$, such that $Lip(f_n) \le 1$ and

$$\int_X f_n(x)(\mu-\nu)(dx) \to W(\mu,\nu).$$

If $\|\mu\| = \|\nu\|$ then for any constant $C \in \mathbb{R}$ it follows that $\int_X C(\mu-\nu)(dx) = 0$. Thus,

$$\infty > K\left(\|\mu\| + \|\nu\|\right) \ge \int_X \|x\|_X |\mu-\nu|(dx) \ge \int_X f_n(x) - f_n(0)(\mu-\nu)(dx) \to W(\mu,\nu).$$

Conversely, if $W(\mu,\nu) < \infty$, then $\int_X C(\mu-\nu)(dx) = 0$ for any $C \in \mathbb{R}$. Consequently, $(\mu-\nu)(\mathbb{R}) = 0$. □

Example 10. Metrics on $\mathfrak{M}(X)$ defined in this section are inherently different from standard metrics on L^p spaces, even if considered on the space of absolutely continuous measures. Consider the following two examples:

1. Let η_ε be a standard mollifier and let $\mu_n,\nu_n \in C^\infty(\mathbb{R})$ be defined as $\mu_n = \frac{1}{n}\delta_0 * \eta_\varepsilon$ and $\nu_n = \frac{1}{n}\delta_{n^2} * \eta_\varepsilon$. For a fixed $\varepsilon > 0$ we have $\|\mu_n - \nu_n\|_{L^p(\mathbb{R})} \to 0$ for any $p \in [1,\infty]$, but also $\|\mu_n - \nu_n\|_W \to \infty$.

2. Let $\mu_n,\nu_n \in C^\infty(\mathbb{R})$ be defined as $\mu_n = n\delta_0 * \eta_{2^{-n}}$ and $v_n = n\delta_{\frac{1}{n^2}} * \eta_{2^{-n}}$, then $\|\mu_n - \nu_n\|_{L^p(\mathbb{R})} \to \infty$ for any $p \in [1,\infty]$, but also $\|\mu_n - \nu_n\|_W \to 0$.

1.1.1. 1-Wasserstein distance

The following characterization of $W(\mu,\nu)$ was derived in [75] for the the case $\mu,\nu\in\mathfrak{M}(\mathbb{R})$.

Theorem 11. 1-*Wasserstein distance between measures* μ *and* ν *on* $\mathbb{R}$ *equals*

$$W(\mu,\nu)=\int_{-\infty}^{\infty}|\mu[-\infty,x]-\nu[-\infty,x]|dx.$$

In other words $W(\mu,\nu)$ *is the* $L^1(\mathbb{R})$ *distance between cumulative distribution functions for* μ *and* ν.

From the definition of $W(\mu,\nu)$ the following propositions follow.

Proposition 12. 1-*Wasserstein distance is scale-invariant, namely*

$$W(\lambda\cdot\mu,\lambda\cdot\nu)=\lambda W(\mu,\nu).$$

Definition 13. Let $x\in X$ and $\mu\in\mathfrak{M}^+(X)$. Define translation of μ by x as

$$T_x\mu(E)=\mu(E+\{-x\}).$$

Proposition 14. 1-*Wasserstein distance is translation-invariant, namely*

$$W(T_x\mu,T_x\nu)=W(\mu,\nu).$$

1.1.2. Normalized Wasserstein distance

By Proposition 9 the 1-Wasserstein distance is not a suitable tool for comparing two measures of different masses. It may seem that the simplest solution is to normalize the measures beforehand. It turns out, however, that $W\left(\frac{\mu}{\|\mu\|},\frac{\nu}{\|\nu\|}\right)$ is not a metric. Instead, the following concept, used for example in [61], may be applied.

Definition 15. We define normalized 1-Wasserstein distance between two measures $\mu,\nu\in\mathfrak{M}(X)$ as

$$\widetilde{W}(\mu,\nu)=\min\left(\|\mu\|+\|\nu\|\,,\,|\,\|\mu\|-\|\nu\|\,|+W\left(\frac{\mu}{\|\mu\|},\frac{\nu}{\|\nu\|}\right)\right). \tag{1.1}$$

Lemma 16. *The distance defined by (1.1) is a metric.*

Proof. Let μ, ν and η be Radon measures. Then, it holds

- $\widetilde{W}(\mu,\nu)=0$ if and only if $\mu=\nu$. Indeed, either $\|\mu\|+\|\nu\|=0$ or $|\,\|\mu\|-\|\nu\|\,|+W\left(\frac{\mu}{\|\mu\|},\frac{\nu}{\|\nu\|}\right)=0$ imply that $\mu=\nu$.
- $\widetilde{W}(\mu,\nu)=\widetilde{W}(\nu,\mu)$,

- Since

$$\begin{aligned}\widetilde{W}(\mu,\nu)+\widetilde{W}(\nu,\eta) &= \min\left(\|\mu\|+\|\nu\|, |\,\|\mu\|-\|\nu\|\,| + W\left(\frac{\mu}{\|\mu\|},\frac{\nu}{\|\nu\|}\right)\right)\\ &\quad + \min\left(\|\eta\|+\|\nu\|, |\,\|\eta\|-\|\nu\|\,| + W\left(\frac{\eta}{\|\eta\|},\frac{\nu}{\|\nu\|}\right)\right),\end{aligned}$$

to show the triangle inequality, we consider four possibilities

$$\begin{aligned}\widetilde{W}(\mu,\nu)+\widetilde{W}(\nu,\eta) &= \|\mu\|+\|\nu\|+\|\eta\|+\|\nu\| \geqslant \|\mu\|+\|\eta\| \geqslant \widetilde{W}(\mu,\eta),\\ \widetilde{W}(\mu,\nu)+\widetilde{W}(\nu,\eta) &= |\,\|\mu\|-\|\nu\|\,| + W\left(\frac{\mu}{\|\mu\|},\frac{\nu}{\|\nu\|}\right)\\ &\quad +|\,\|\eta\|-\|\nu\|\,| + W\left(\frac{\eta}{\|\eta\|},\frac{\nu}{\|\nu\|}\right) \geqslant \widetilde{W}(\mu,\eta),\\ \widetilde{W}(\mu,\nu)+\widetilde{W}(\nu,\eta) &= \|\mu\|+\|\nu\|+|\,\|\eta\|-\|\nu\|\,| + W\left(\frac{\eta}{\|\eta\|},\frac{\nu}{\|\nu\|}\right) \geqslant \|\mu\|+\|\eta\|\\ &\geqslant \widetilde{W}(\mu,\eta),\\ \widetilde{W}(\mu,\nu)+\widetilde{W}(\nu,\eta) &= \|\eta\|+\|\nu\|+|\,\|\mu\|-\|\nu\|\,| + W\left(\frac{\mu}{\|\mu\|},\frac{\nu}{\|\nu\|}\right) \geqslant \|\mu\|+\|\eta\|\\ &\geqslant \widetilde{W}(\mu,\nu).\end{aligned}$$

□

This metric lacks the scaling property (namely in general $\widetilde{W}(\lambda\mu,\lambda\nu)=\lambda\widetilde{W}(\mu,\nu)$ does not hold). Nonetheless, the following weaker property holds.

Proposition 17. *Let μ_k and ν_k be two sequences of Radon measures and $\|\mu_k\|\to 0$, $\|\nu_k\|\to 0$ then $\widetilde{W}(\mu_k,\nu_k)\to 0$.*

Note that $W\left(\frac{\mu_k}{\|\mu_k\|},\frac{\nu_k}{\|\nu_k\|}\right)$ does not satisfy the weaker property, since for $\mu_k=\frac{1}{k}\delta_0$, $\nu_k=\frac{1}{k}\delta_1$ we have

$$\lim_{k\to\infty} W\left(\frac{\frac{1}{k}\delta_0}{\left\|\frac{1}{k}\delta_0\right\|},\frac{\frac{1}{k}\delta_1}{\left\|\frac{1}{k}\delta_1\right\|}\right)=1.$$

1.1.3. Centralized Wasserstein metric

For applications that require scale-invariance and comparing measures of unequal masses neither 1-Wasserstein nor Normalized Wasserstein distance is suitable.

Definition 18. Centralized 1-Wasserstein distance between two measures $\mu,\nu\in\mathfrak{M}(X)$ reads

$$\widehat{W}(\mu,\nu)=\sup\left\{\int_X f d(\mu-\nu):\ f\in C^{0,1}(X),\ Lip(f)\leq 1,\ |f(0)|\leq 1\right\}.$$

This metric was introduced in [36] for analysis of the measure-valued structured population models.

This metric is scale-invariant, but in contrast to Wasserstein metric, it is not translation-invariant. Applications of centralized 1-Wasserstein metric are therefore restricted to modeling of specific phenomena, for which the dependence of error on location in X is justifiable.

Consider the following example: $\mu_x = 2\delta_x$, $\nu_x = 3\delta_x$. If measures μ and ν represent structure distribution of a population (e.g. μ_x is a model prediction of size-distribution of a population and ν_x is an empirical size-distribution computed based on experimental data) and moreover new individuals are always born with a fixed structural variable $x_0 \in X$ one may argue that the error, $e(\mu_x, \nu_x)$, should depend on x. The difference of masses at $x \in X$ is a result of both the difference in the number newborns (with structural variable x_0) and the individual growth process from x_0 to x. Consequently, one would expect that for two structural points $x, y \in X$ condition $\|x - x_0\|_X \geq \|y - x_0\|_X$ implies $e(\mu_x, \nu_x) \geq e(\mu_y, \nu_y)$. Centralized Wasserstein metric meets this expectation since in that case $\widehat{W}(\mu_x, \nu_x) \geq \widehat{W}(\mu_y, \nu_y)$. On the other hand, the above argumentation is hard to defend if mortality, and therefore mass annihilation at every point of X, is involved. In the next section, a more versatile and translation-invariant metric is introduced.

1.1.4. Bounded Lipschitz distance

The flat metric, known also as a bounded Lipschitz distance [60], is scale- and translation-invariant. It has proven to be useful in analysis of structured population models and, in particular, Lipschitz dependence of solutions on the model parameters and initial data [35, 11]. The flat metric has been recently used for the proof of convergence and stability of EBT numerical scheme (see [9, 11]).

The following three lemmas provide tools for estimating ρ_F from above. The first estimate arises from Proposition 7 and its proof can be found in Section 7 in [34].

Lemma 19. *Let $\mu, \nu \in \mathfrak{M}_d^+(X)$ and $\mu = \sum_{i=1}^N m_i \delta_{x_i}$, $\nu = \sum_{i=1}^N n_i \delta_{y_i}$ then*

$$\rho_F(\mu, \nu) \leq \sum_{i=1}^N |m_i - n_i| + \sum_{i=1}^N \|x_i - y_i\|_X \, n_i.$$

Proof. Let $\tilde{\mu} = \sum_{i=1}^N n_i \delta_{x_i}$. From triangle inequality we obtain

$$\rho_F(\mu, \nu) \leq \rho_F(\mu, \tilde{\mu}) + \rho_F(\tilde{\mu}, \nu) = \|\mu - \tilde{\mu}\| + W(\tilde{\mu}, \nu).$$

Directly from the definitions of appropriate metrics it follows that

$$\|\mu - \tilde{\mu}\| = \sum_{i=1}^N |m_i - n_i|$$

and

$$W(\tilde{\mu}, \nu) = \sum_{i=1}^N n_i \left(f(x_i) - f(y_i)\right),$$

for some $f \in C^{0,1}(X)$ satisfying $|f(x_i) - f(y_i)| \leq \|x_i - y_i\|_X$. Finally, we obtain

$$W(\tilde{\mu}, \nu) \leq \sum_{i=1}^{N} \|x_i - y_i\|_X \, n_i,$$

which completes the proof. □

The second lemma is a straightforward corollary resulting from definition of flat distance.

Lemma 20. *Let μ and ν be two non-negative Radon measures on $X = X_1 \cup X_2$ with $X_1 \cap X_2 = \emptyset$. Then*

$$\rho_F(\mu, \nu) \leq \rho_F(\mu|_{X_1}, \nu|_{X_1}) + \rho_F(\mu|_{X_2}, \nu|_{X_2})$$

The following fact follows directly from the definition of flat distance.

Lemma 21. *For every $\mu, \nu \in \mathfrak{M}(X)$ and $f \in C^{0,1}(X)$ it holds that*

$$\rho_F(\mu, \tilde{\mu}) \geq \frac{\int_X f d\,(\mu - \tilde{\mu})}{\|f\|_{C_b^{0,1}(X)}}.$$

An easy, yet important, conclusion from Lemma 19 can be made.

Corollary 22. *For $\mu, \nu \in \mathfrak{M}_d(X)$ we have*

$$\rho_F(\mu, \nu) \leq \inf_{\substack{\|\tilde{\mu}\| = \|\nu\| \\ \tilde{\mu} \in \mathfrak{M}_d(X)}} \|\mu - \tilde{\mu}\| + W(\tilde{\mu}, \nu).$$

Proof. Lemma 19 can be reformulated as $\rho_F(\mu, \nu) \leq \|\mu - \tilde{\mu}\| + W(\tilde{\mu}, \nu)$ for any $\mu, \nu \in \mathfrak{M}_d(X)$ and $\tilde{\mu}$ being supported on a subset of $supp\mu \cup supp\nu$ with $\|\tilde{\mu}\| = \|\nu\|$. Since there are no assumptions on N, and also m_i, n_i are not necessarily strictly positive measure $\tilde{\mu}$ can be supported on an arbitrary discrete subset of X. □

A farther generalization of Corollary 22 is provided by Theorem 25.

1.2. Dual representation

The following two theorems connect Wasserstein metric with transportation theory and provide a dual representation for $W(\mu, \nu)$. Proofs can be found in [76].

Theorem 23. *(Kantorovich and Rubinstein) Wasserstein distance between probability measures μ and ν on a metric space (X, d) equals*

$$W(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu, \nu)} \left\{ \int_{X \times X} d(x, y) d\gamma \right\}$$

where $\Gamma(\mu, \nu)$ denotes a subset of $\mathfrak{M}^+(X \times X)$ of all measures with marginals equal to μ and ν on the first and second factors respectively. $\Gamma(\mu, \nu)$ is often referred to as the set of transference plans.

Theorem 24. *For every pair of measures μ and ν on a metric space (X, d) there exists an optimal transference plan γ^* such that*

$$W(\mu, \nu) = \int_{X \times X} d(x, y)_X d\gamma^*$$

An analogue of Theorem 23 for flat metric was first noticed in [40] and proved in [62] for the case of $X = \mathbb{R}^d$.

Theorem 25. *Bounded Lipschitz distance between finite Radon measures μ and ν on $\mathbb{R}^d$ equals*

$$\inf_{\substack{\tilde{\mu}, \tilde{\nu} \in \mathfrak{M}(\mathbb{R}^d) \\ \|\tilde{\mu}\| = \|\tilde{\nu}\|}} \|\mu - \tilde{\mu}\| + \|\nu - \tilde{\nu}\| + W(\tilde{\mu}, \tilde{\nu}).$$

In fact, intermediate measures $\tilde{\mu}$ and $\tilde{\nu}$ for which the infimum is attained are always no greater than μ and ν respectively. The following result was proved in Section 2.1 in [62]:

Corollary 26. *Let $\mu, \nu \in \mathfrak{M}(\mathbb{R}^d)$ then*

$$\rho_F(\mu, \nu) = \inf_{\substack{\tilde{\mu}, \tilde{\nu} \in \mathfrak{M}(\mathbb{R}^d) \\ \tilde{\mu} \leq \mu, \tilde{\nu} \leq \nu \\ \|\tilde{\mu}\| = \|\tilde{\nu}\|}} \|\mu - \tilde{\mu}\| + \|\nu - \tilde{\nu}\| + W(\tilde{\mu}, \tilde{\nu}).$$

Dual representations allows for easier reasoning about upper bounds of distances. For instance Corollary 22, which generalizes Lemma 19 follows immediately from dual representation of flat metric. Similarly does Theorem 11. Another profit arising from the dual representations is that an approach based on flow networks can be used to compute the value of the distance (see Section 1.3.1).

1.3. Computational complexity

In this section algorithmic aspects of numerical computation of distances between two non-negative discrete Radon measures are discussed. The set of discrete measures, $\mathfrak{M}_d(X)$, is dense in $\mathfrak{M}(X)$ hence the distance between arbitrary two measures can be computed by approximating each of them with a discrete measure (see Theorem 41).

Each of the considered distances can be determined by linear programming. Computational complexity of this approach is often too large for applications. For the case of arbitrary space X we present how the problem can be reduced to finding a maximum-flow minimum-cost for a bipartite graph. For the case of $X = \mathbb{R}$, Theorem 11 provides an alternative approach which leads to a linear algorithm for 1-Wasserstein distance. Moreover, an analogue of Theorem 11 is presented and an algorithm for computing flat metric is derived.

Unless stated otherwise, by the input length of a problem, N, we mean the number of Dirac masses in both of the compared measures. The aim of this section is to present efficient algorithms for Wasserstein-type metrics described in Section 1.1. In particular, a novel algorithm for computing the flat metric on $\mathbb{R}$ with computational cost $O(N \log N)$ is proposed.

1.3.1. Transference plan as a flow network

Given two discrete measures $\mu, \nu \in \mathfrak{M}_{d,N}(X)$ the problem of computing $W(\mu,\nu)$, $\widehat{W}(\mu,\nu)$ and $\rho_F(\mu,\nu)$ can be reduced to an instance of linear programming. Indeed, let

$$\mu - \nu = \sum_{i=1}^{N} m_i \delta_{x_i},$$

then $W(\mu,\nu)$ maximizes linear objective function

$$c(f_1, f_2, ..., f_N) = \sum_{i=1}^{N} m_i f_i$$

subject to the following linear inequality constraints:

$$\begin{aligned} f_i - f_{i+1} &\leq x_{i+1} - x_i \\ f_{i+1} - f_i &\leq x_{i+1} - x_i \end{aligned}$$

for every $i \in \{1, 2, ..., N-1\}$. Similarly, the distance $\widehat{W}(\mu,\nu)$ maximizes the same objective function, c, subject to additional constraint given by

$$\begin{aligned} f_{i+1} &\leq 1 + x_{i+1} - x_i \\ f_{i+1} &\geq -1 - (x_{i+1} - x_i) \\ f_{i-1} &\leq 1 + x_i - x_{i-1} \\ f_{i-1} &\geq -1 - (x_i - x_{i-1}) \end{aligned}$$

for $x_{i-1} < 0 < x_{i+1}$. Finally, flat distance $\rho_F(\mu,\nu)$ also maximizes c and requires additional constraints given by

$$\begin{aligned} f(x_i) &\leq 1 \\ f(x_i) &\geq -1 \end{aligned}$$

for every $i \in \{1, 2, ..., N\}$.

Despite the fact that linear programming has been studied intensively since the beginning of 20^{th} century, a question whether there exists a sub-exponential algorithm solving the linear programming problem remained open until 1979. The current opinion is that the efficiency of good implementations of exponential simplex-based methods and polynomial interior point methods are similar [31]. In this section we present a method of reducing the problem of computing $W(\mu,\nu)$ to an instance of a maximum-flow minimum-cost problem. It is beneficial since efficient algorithms for solving this problem for the case of bipartite graphs are known [58, 19]. Finally, a generalization of this method, inspired by [45], to the case of flat metric, $\rho_F(\mu,\nu)$ is presented.

Definition 27. Flow network is a finite directed graph (V, E) with a capacity function $w : V \times V \to \mathbb{R} \cup \{\infty\}$ and a cost function $c : V \times V \to \mathbb{R}$.

In this section we show that Wasserstein distance between two discrete probabilistic measures and bounded Lipschitz distance between two discrete Radon measures on a metric space X, with a finite number of atoms ($\mu = \sum_{i=1}^{N} m_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{M} n_j \delta_{y_j}$) can be computed using maximum-flow minimum-cost approach.

Definition 28. For given probabilistic measures $\mu = \sum_{i=1}^{N} m_i \delta_{x_i}$, $\nu = \sum_{j=1}^{M} n_j \delta_{y_j}$ we define a Wasserstein flow network $N_W = (V_W, E_W)$ by

$$\begin{aligned} V_W &= \{s, x_1, x_2, ..., x_N, y_1, y_2, ..., y_M, t\} \\ E_W &= \{s\} \times \{x_1, ..., x_N\} \cup \{y_1, ..., y_M\} \times \{t\} \cup \{x_1, ..., x_N\} \times \{y_1, ..., y_M\} \end{aligned}$$

with a capacity function

$$w(u,v) = \begin{cases} m_i & \text{if } u = s \text{ and } v = x_i \\ n_i & \text{if } u = y_i \text{ and } v = t \\ \infty & \text{otherwise} \end{cases}$$

and cost function

$$c(u,v) = \begin{cases} d(x_i, y_i) & \text{if } u = x_i \text{ and } v = y_i \\ 0 & \text{otherwise} \end{cases}$$

Network N_W is depicted on Figure 1.1.

Definition 29. A flow in a flow network $N = (V, E)$ is a mapping $f : E \to \mathbb{R}^{\geq 0}$, subject to the following constraints:

1. for every $(u, v) \in E$ it holds that $f(u, v) \leq w(u, v)$, where w is the capacity function
2. for every $v \in V \setminus \{s, t\}$ it holds that $\sum_{\{u:(u,v)\in E\}} f(u, v) = \sum_{\{u:(v,u)\in E\}} f(v, u)$

Definition 30. A maximum-flow in a flow network $N = (V, E)$ is a flow, f, that maximizes $\sum_{\{v:(s,v)\in E\}} f(s, v)$.

Definition 31. A maximum-flow minimum-cost is the minimal value of

$$\sum_{(u,v)\in E} c(u,v) f(u,v)$$

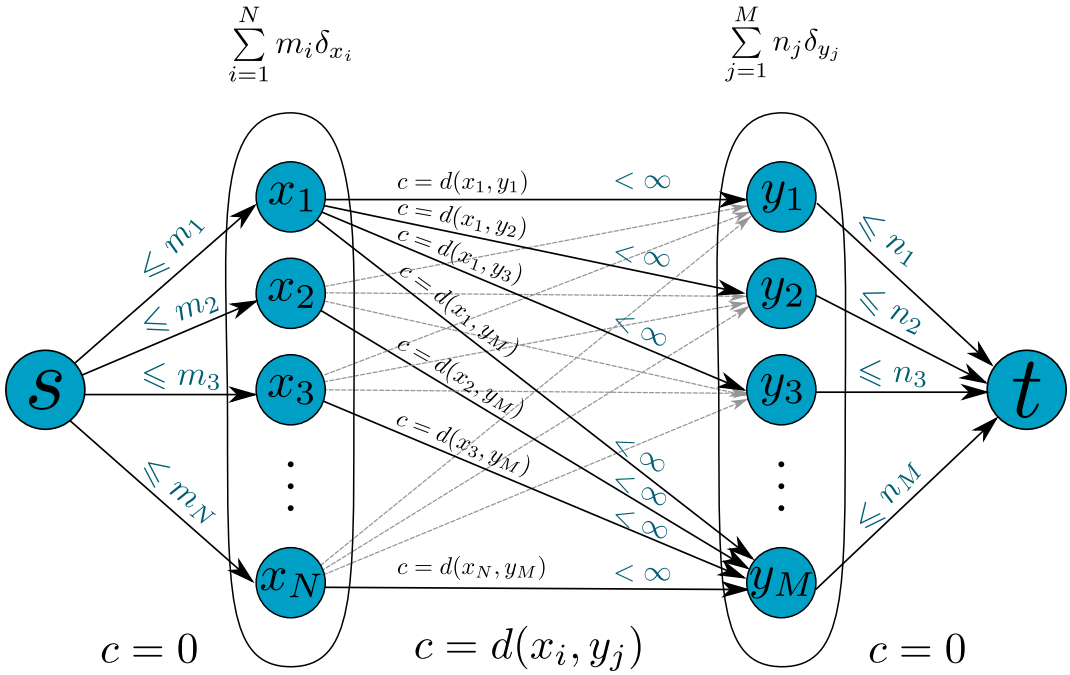
for f being a maximum flow.

Theorem 32. *The maximum-flow minimum-cost of network N_W equals $W(\mu, \nu)$.*

Proof. Every transference plan $\gamma \in \Gamma(\mu, \nu)$ defines a maximum flow in network N_W by

$$f_\gamma(u,v) = \begin{cases} w(u,v) & \text{if } u = s \text{ or } v = t \\ \gamma(\{u\}, \{v\}) & \text{otherwise} \end{cases},$$

Figure 1.1: Wasserstein flow network N_W for measures $\sum_{i=1}^{N} m_i \delta_{x_i}$ and $\sum_{j=1}^{M} n_j y_j$.



and every maximum flow defines a transference plan. Moreover the cost associated with flow f_γ equals

$$\sum_{u \in \{x_1,...,x_N\}} \sum_{v \in \{y_1,...,y_M\}} c(u,v) f_\gamma(u,v) = \int_{X \times X} d(x,y) d\gamma$$

Consequently, by Theorem 23, the maximum-flow minimum-cost of network N_W equals $W(\mu, \nu)$. □

Definition 33. For given measures $\mu = \sum_{i=1}^{N} m_i \delta_{x_i}$, $\nu = \sum_{j=1}^{M} n_j \delta_{y_j}$ we define a flat flow network $N_F = (V_F, E_F)$ by

$$\begin{aligned} V_F &= \{s, x_1, x_2, ..., x_N, y_1, y_2, ..., y_M, t\} \\ E_F &= \{s\} \times (V_F \setminus \{s,t\}) \cup (V_F \setminus \{s,t\}) \times \{t\} \cup \{x_1, ..., x_N\} \times \{y_1, ..., y_M\} \end{aligned}$$

with a capacity function

$$w(u,v) = \begin{cases} m_i & \text{if } u = s \text{ and } v = x_i \\ n_i & \text{if } u = y_i \text{ and } v = t \\ \infty & \text{otherwise} \end{cases}$$

and cost function

$$c(u,v) = \begin{cases} d(x_i, y_i) & \text{if } u = x_i \text{ and } v = y_i \\ 1 & \text{if } u = x_i \text{ and } v = t \\ 1 & \text{if } u = s \text{ and } v = y_i \\ 0 & \text{otherwise} \end{cases}$$

Network N_W is depicted on Figure 1.2.

Theorem 34. *The maximum flow minimum cost of network N_F equals $\rho_F(\mu,\nu)$.*

Proof. Every choice of $(\tilde{\mu},\tilde{\nu},\gamma) \in \mathfrak{M}_d(X)\times\mathfrak{M}_d(X)\times\Gamma(\tilde{\mu},\tilde{\nu})$ such that $\sum_{i=1}^{N}\tilde{m}_i\delta_{x_i} = \tilde{\mu} \leq \mu$ and $\sum_{i=1}^{M}\tilde{n}_i\delta_{y_i} = \tilde{\nu} \leq \nu$ defines a maximum flow in network N_F by

$$f_{\tilde{\mu},\tilde{\nu},\gamma}(u,v) = \begin{cases} w(u,v) & \text{if } u = s \text{ and } v = x_i \\ w(u,v) & \text{if } u = y_i \text{ and } v = t \\ m_i - \tilde{m}_i & \text{if } u = x_i \text{ and } v = t \\ n_i - \tilde{n}_i & \text{if } u = s \text{ and } v = y_i \\ \gamma(\{u\},\{v\}) & \text{otherwise} \end{cases},$$

and every maximum flow defines a triple $(\tilde{\mu},\tilde{\nu},\gamma) \in \mathfrak{M}_d(X)\times\mathfrak{M}_d(X)\times\Gamma(\tilde{\mu},\tilde{\nu})$. Moreover the cost associated with a flow $f_{\tilde{\mu},\tilde{\nu},\gamma}$ equals

$$\sum_{u\in\{x_1,\ldots,x_N,s\}}\sum_{v\in\{y_1,\ldots,y_M,t\}} c(u,v)f_{\tilde{\mu},\tilde{\nu},\gamma}(u,v) =$$

$$= \int_{X\times X} d(x,y)d\gamma + \sum_{i=1}^{N}(m_i - \tilde{m}_i) + \sum_{i=1}^{M}(n_i - \tilde{n}_i)$$

Consequently, by Corollary 26, the maximum flow minimum cost of network N_F equals $\rho_F(\mu,\nu)$. □

Since network N is a bipartite graph (excluding s and t vertices) the Hungarian algorithm [58, 45, 19] can be applied to compute 1-Wasserstein and Bounded Lipschitz distances. This approach proves to be significantly more efficient than general linear programming.

Example 35. Let us consider the following expression: $\rho_F(2\delta_x, 3\delta_y)$ for some $x, y \in X$. The value of this distance can be computed by following methods:

- Let $N_F = (V,E)$ be the flat flow network for measures $2\delta_x$ and $3\delta_y$, hence $V = \{s,x,y,t\}$. By definition the maximum-flow in N_F is a flow, f, which maximizes $f(s,x) + f(s,y)$. Since $f(s,x) \leq w(s,x) = 2$ and $f(y,t) \leq w(y,t) = 3$ we infer that for the maximum-flow $f(s,x) = 2$ and $f(y,t) = 3$. Since $f(s,x) = f(x,y) + f(x,t)$ and $f(x,y) + f(s,y) = f(y,t)$, we conclude that the cost of a maximum-flow equals

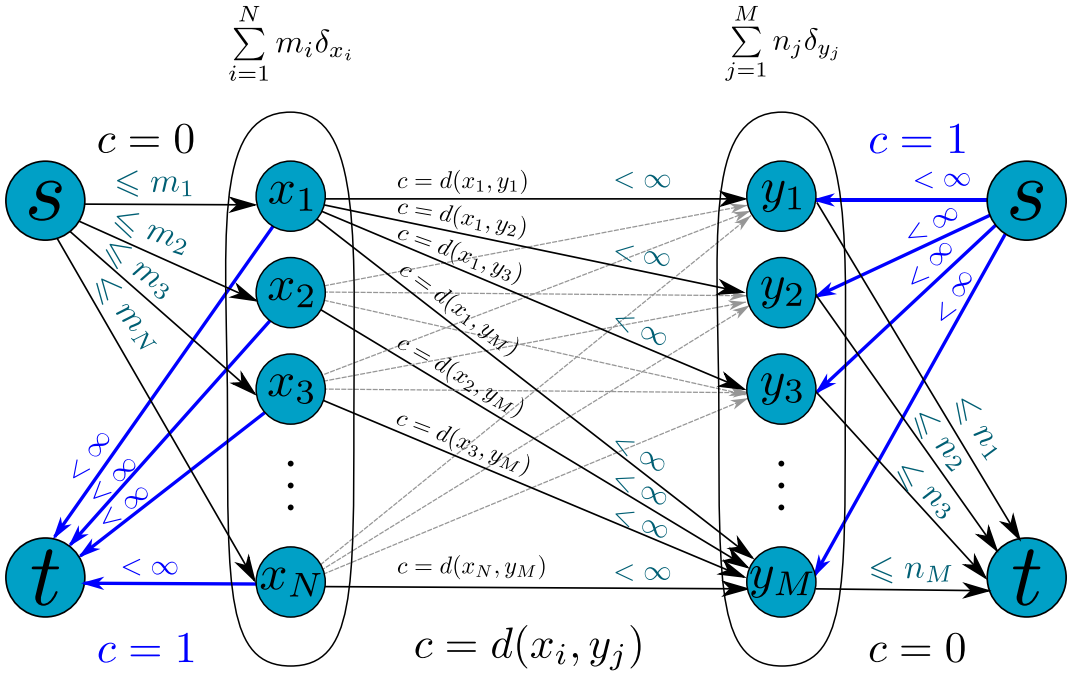
$$f(x,t) + d(x,y)f(x,y) + f(s,y) =$$

$$= (2+3) + (d(x,y) - 2)\, f(x,y).$$

  It is easy to check that for any value $f(x,y) \in [0,2]$ a maximum-flow can be built. Finally, by Theorem 34 we obtain

$$\rho_F(2\delta_x, 3\delta_y) = \begin{cases} 5 & \text{if } d(x,y) \geq 2 \\ 1 + 2d(x,y) & \text{otherwise} \end{cases}.$$

Figure 1.2: Flat flow network N_W for measures $\sum_{i=1}^{N} m_i\delta_{x_i}$ and $\sum_{j=1}^{M} n_j y_j$. Edges from the set $E_F \setminus E_W$ are shown in blue.



- By the definition of $\rho_F(\mu, \nu)$ we have that

$$\rho_F(2\delta_x, 3\delta_y) = \sup\{2f_x - 3f_y : |f_x - f_y| \leq d(x,y),\ |f_x| \leq 1,\ |f_y| \leq 1\}.$$

Fix $f_x \in [-1, 1]$, then

$$f_y \in [f_x - d(x,y), f_x + d(x,y)] \cap [-1, 1].$$

Since $-3f_y$ is decreasing with f_y it attains maximum value at

$$f_y = \max(f_x - d(x,y), -1).$$

Consequently,

$$\begin{aligned}\rho_F(2\delta_x, 3\delta_y) &= \sup_{f_x \in [-1,1]} \{2f_x - 3\max(f_x - d(x,y), -1)\} = \\ &= \sup_{f_x \in [-1,1]} \min(3d(x,y) - f_x, 3 + 2f_x).\end{aligned}$$

If $d(x,y) \geq 2$ then $3 + 2f_x \leq 3d(x,y) - f_x$ for every $f_x \in [-1, 1]$, thus

$$\rho_F(2\delta_x, 3\delta_y) = \sup_{f_x \in [-1,1]} 3 + 2f_x = 5.$$

Otherwise, if $d(x,y) < 2$ then $3 + 2f_x = 3d(x,y) - f_x$ for $f_x = d(x,y) - 1$. Since $3 + 2f_x$ is increasing and $3d(x,y) - f_x$ is decreasing we obtain

$$\rho_F(2\delta_x, 3\delta_y) = 3d(x,y) - (d(x,y) - 1) = 1 + 2d(x,y).$$

1.3.2. 1-Wasserstein distance on $\mathfrak{M}_d^+(\mathbb{R})$

Theorem 11 provides tools for computing Wasserstein distance as an integral which in the case of discrete measures is simply a finite sum of N elements. In this section we derive this algorithm again, from a different perspective in a seemingly overcomplicated way. The purpose of this is to make an introduction to this approach, which is farther applied for more involved algorithms for other distances.

Let $\mu, \nu \in \mathfrak{M}_d^+(\mathbb{R})$, $\|\mu\| = \|\nu\|$, and $\mu - \nu = \sum_{k=1}^N m_k \delta_{x_k}$. Since $\int_X C d(\mu - \nu) = 0$ we can add an arbitrary constant to the test function in the definition of 1-Wasserstein distance and hence

$$W(\mu, \nu) = \sup \left\{ \sum_{k=1}^N m_k f(x_k) : f \in C(\mathbb{R}), f(x_N) = 0, Lip(f) \leqslant 1 \right\}.$$

Regularity conditions can be represented as linear programming bounds. Hence, computing of $W(\mu, \nu)$ is equivalent to finding maximum of

$$\left| \sum_{k=1}^N m_k f_k \right|$$

with the following restrictions

$$\begin{aligned} f_N &= 0, \\ |f_k - f_{k-1}| &\leqslant |x_k - x_{k-1}|. \end{aligned}$$

Although this problem can clearly be solved by linear programming, a more efficient algorithm can be found. Define

$$W^m(f) = \sup \left\{ \sum_{k=1}^m m_k f_k : \{f_i\}_{i=0}^N \subset \mathbb{R}, f_m = f, \forall_{k \in \{1,\ldots,N\}} |f_k - f_{k-1}| \leqslant |x_k - x_{k-1}| \right\}.$$

Obviously $W(\mu, \nu) = W^N(0)$. Denote $d_k = x_{k+1} - x_k$, and observe that the value of $W^m(f)$ can be computed recursively as follows

$$W^1(f) = m_1 x,$$

$$\begin{aligned} W^2(f) &= m_2 f + \sup_{f_1 \in [f - d_1, f + d_1]} W^1(f) = m_2 f + m_1 f + m_1 \cdot sgn(m_1) d_1 = \\ &= (m_1 + m_2) x + |m_1| d_1. \end{aligned}$$

It can be shown by induction that

$$W^N(f) = \left(\sum_{i=1}^N m_i \right) f + \sum_{i=1}^{N-1} d_i \left| \sum_{j=1}^i m_j \right|. \tag{1.2}$$

Notice that the value m_N is not used in the formula for $W^N(0)$. It is, however, involved indirectly, because $m_N = -\sum_{i=1}^{N-1} m_i$.

1.3.2.1. Pseudocode

Equation (1.2) gives an explicit formula for $W(\mu,\nu)$, which is trivial to compute. Nonetheless, in this section we provide a pseudocode for computing iterated sum $\sum_{i=1}^{N-1} d_i \left|\sum_{j=1}^{i} m_j\right|$ in linear time to make sure the reader is familiar with pseudocode notation before moving forward to more involved examples.

In this algorithm we initially assign 0 value to variables '*distance*' and '*partialSum*' and then process the array of positions, x, and the array of masses, m, sequentially. In each iteration one, consecutive, index idx is processed. After indices $\{1,2,3,...,k\}$ were processed the variable $partialSum$ contains $\sum_{j=1}^{k} m_j$ and $distance$ contains $\sum_{i=1}^{k} d_i \left|\sum_{j=1}^{i} m_j\right|$. Consequently, after all indices smaller than N are processed the returned variable $distance$ contains $W(\mu,\nu)$.

`Input:`

- `non-decreasing table of positions,` $x \in \mathbb{R}^N$,
- `table of masses,` $m \in \mathbb{R}^N$.

`1-`WASSERSTEIN-DISTANCE `(`$x \in \mathbb{R}^N$`,` $m \in \mathbb{R}^N$`):`
$distance \leftarrow 0$
$partialSum \leftarrow 0$
`for` $idx \leftarrow 1$ `to` $N-1$ `do`

$partialSum \leftarrow partialSum + m_{idx}$
$distance \leftarrow distance + (x_{idx+1} - x_{idx}) \cdot |partialSum|$

`return` $distance$

1.3.2.2. Complexity of the algorithm

It is clear from the pseudocode that the computational complexity of the algorithm is $\Theta(N)$, while memory complexity (the volume of memory used by the algorithm) is $\Theta(1)$.

1.3.3. Centralized Wasserstein distance on $\mathfrak{M}_d^+(\mathbb{R})$

Let

$$\mu - \nu = \sum_{i=1}^{M} m_i \delta_{x_i} + m_{M+1}\delta_0 + \sum_{i=M+2}^{N} m_i \delta_{x_i}.$$

Define

$$\underline{W}^j(f) = \sup\left\{\sum_{k=1}^{j} m_k f_k : \{f_i\}_{i=0}^{N} \subset \mathbb{R}, f_j = f, \forall_{k\in\{1,..,j\}} |f_k - f_{k-1}| \leqslant |x_k - x_{k-1}|\right\},$$

$$\overline{W}^j(f) = \sup\left\{\sum_{k=j}^{N} m_k f_k : \{f_i\}_{i=0}^{N} \subset \mathbb{R}, f_j = f, \forall_{k\in\{1,..,j\}} |f_k - f_{k-1}| \leqslant |x_k - x_{k-1}|\right\}.$$

As already proven

$$\begin{aligned}\underline{W}^{M+1}(f) &= \left(\sum_{i=1}^{M+1} m_i\right) f + \sum_{k=1}^{M} d_k \left|\sum_{i=1}^{k} m_i\right|, \\ \overline{W}^{M+1}(f) &= \left(\sum_{i=M+1}^{N} m_i\right) f + \sum_{k=1}^{N-(M+1)} d_{N-k} \left|\sum_{i=N+1-k}^{N} m_i\right|.\end{aligned}$$

From the definition it can be deduced that

$$\widehat{W}(\mu,\nu) = \sup_{f\in[-1,1]} \left(\underline{W}^{M+1}(f) + \overline{W}^{M+1}(f) - m_{M+1} f\right),$$

so the distance is given by the formula

$$\widehat{W}(\mu,\nu) = \sum_{k=1}^{M} d_k \left|\sum_{i=1}^{k} m_i\right| + \sum_{k=1}^{N-(M+1)} d_{N-k} \left|\sum_{i=N+1-k}^{N} m_i\right| + \left|\sum_{i=1}^{N} m_i\right|.$$

1.3.3.1. Pseudocode

Similarly as in the case of 1-Wasserstein distance the algorithm is straightforward. It consists of three loops. In the first two *while* loops terms

$$\sum_{k=1}^{M} d_k \left|\sum_{i=1}^{k} m_i\right|$$

and

$$\sum_{k=1}^{N-(M+1)} d_{N-k} \left|\sum_{i=N+1-k}^{N} m_i\right|$$

are computed exactly as in 1-Wasserstein-Distance. Finally, in the third loop remaining masses (corresponding to position 0) are added to variable $partialSumFront$, to ensure that $partialSumFront + partialSumBack = \sum_{i=1}^{N} m_i$.

`Input`:

- `non-decreasing table of positions`, $x \in \mathbb{R}^N$,
- `table of masses`, $m \in \mathbb{R}^N$.

`Wasserstein-Centralized-Distance`($x \in \mathbb{R}^N$, $m \in \mathbb{R}^N$):
$distance \leftarrow 0$
$(partialSumFront, partialSumBack) \leftarrow (0, 0)$
$(idxFront, idxBack) \leftarrow (1, N)$
`while` $x_{idxFront} < 0$ `do`

$partialSumFront \leftarrow partialSumFront + m_{idxFront}$
$distance \leftarrow distance + (x_{idxFront+1} - x_{idxFront}) \cdot |partialSumFront|$
$idxFront \leftarrow idxFront + 1$

```
while x_{idxBack} > 0 do
    partialSumBack ← partialSumBack + m_{idxEnd}
    distance ← distance + (x_{idxBack} − x_{idxBack−1}) · |partialSumBack|
    idxBack ← idxBack − 1
for idx ← idxFront to idxBack do
    partialSumFront ← partialSumFront + m_{idx}
return distance + |partialSumFront + partialSumBack|
```

1.3.3.2. Complexity of the algorithm

Each iteration of each loop takes a constant time. The total number of iterations in all three loops is equal to $M+1+(N-M-1)$. Computational complexity of this algorithm is therefore $\Theta(N)$, while the memory complexity is $\Theta(1)$.

1.3.4. Flat distance on $\mathfrak{M}_d^+(\mathbb{R})$

In this section the main result from [40], namely the algorithm for computing flat distance in $\mathcal{O}(N \log N)$, is presented.

Computing flat distance requires storing the shape of functions analogous to W^m as they get more complicated when m increases. We provide a recursive formula for the sequence of these functions. The pseudocode in Section 1.3.4.1 implements the algorithm using an abstract data structure, without specifying its exact implementation, to store previously defined functions. However, the computational complexity depends on the particular choice of this structure. In further sections we provide two solutions that require respectively $\mathcal{O}(N^2)$ and $\mathcal{O}(N \log N)$ operations.

Let

$$\mu - \nu = \sum_{i=1}^{N} m_i \delta_{x_i}.$$

Computing of $F(\mu, \nu)$ is equivalent to finding maximum of

$$\left| \sum_{k=1}^{N} m_k f_k \right|$$

with the following restrictions

$$\begin{aligned} |f_k| &\leqslant 1, \\ |f_k - f_{k-1}| &\leqslant |x_k - x_{k-1}|. \end{aligned}$$

Define

$$F^m(f) = \sup \left\{ \sum_{k=1}^{m} m_k f_k : \{f_i\}_{i=0}^{N} \subset [-1,1], f_m = f, \forall_{i \in \{1,..,N\}} |f_k - f_{k-1}| \leqslant |x_k - x_{k-1}| \right\}.$$

By the definition of flat metric

$$F(\mu,\nu) = \sup_{x\in[-1,1]} F^N(x).$$

Observe that

$$\begin{aligned} F^1(f) &= m_1 f, \\ F^2(f) &= m_2 f + \sup_{f_1\in[f-d_1,f+d_1]\cap[-1,1]} F^1(f_1) = m_2 f + \min(|m_1|, m_1 f + |m_1| d_1), \\ \ldots &\ \ldots\ \ldots \end{aligned}$$

$$F^m(f) = m_m f + \sup_{f_{m-1}\in[f-d_{m-1},f+d_{m-1}]\cap[-1,1]} F^{m-1}(f_{m-1}). \tag{1.3}$$

Computing of F^m based on F^{m-1} is more complex than computing W^m based on W^{m-1}, because F^{m-1} is not necessarily monotonic. The following two lemmas and Figure 1.3 explain the relation between F^m and F^{m-1}.

Lemma 36. *Function F^m is concave for each m.*

Proof. To prove the lemma we use induction with respect to m. $F^1(f)$ is given as $a_1 f$, so it is indeed concave. Assume F^m is concave. Define

$$F^{n,d}_{max}(f) = \sup_{y\in[f-d,f+d]\cap[-1,1]} F^n(y).$$

Choose $x, y \in [-1,1]$. Then, there exist $x' \in B(x,d)\cap[-1,1]$, $y' \in B(y,d)\cap[-1,1]$ such that

$$\alpha F^{m,d}_{max}(x) + (1-\alpha) F^{m,d}_{max}(y) = \alpha F^m(x') + (1-\alpha) F^m(y').$$

Because F^m is concave, it holds

$$\alpha F^m(x') + (1-\alpha) F^m(y') \leqslant F^m\left(\alpha x' + (1-\alpha) y'\right) \leqslant F^{m,d}_{max}(\alpha x + (1-\alpha) y)$$

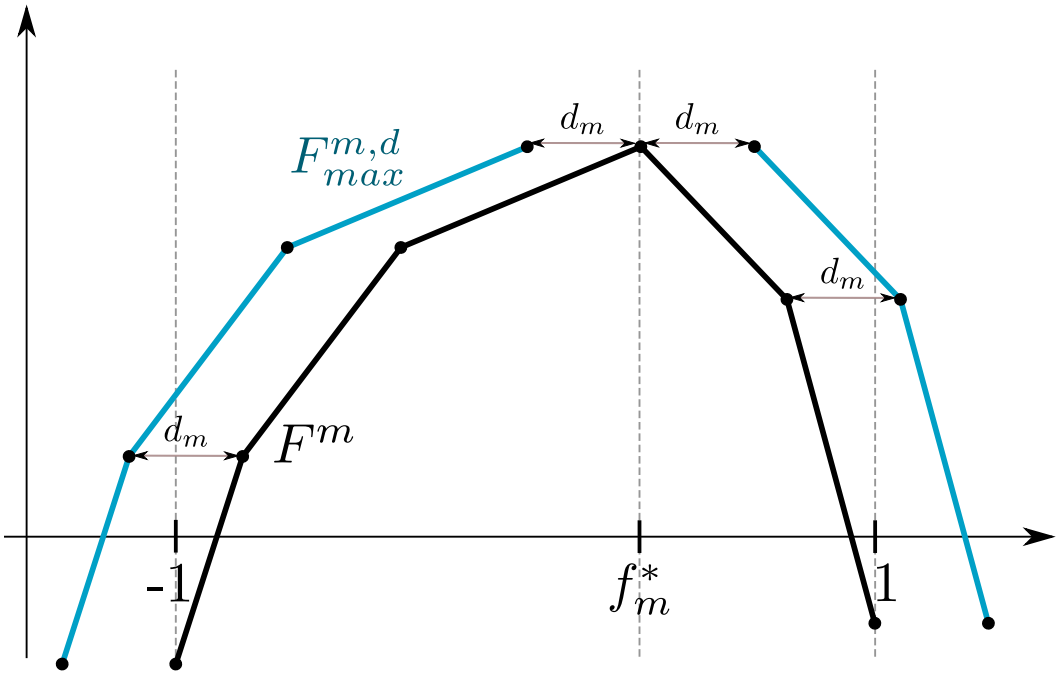
The last inequality follows from $\alpha x' + (1-\alpha) y' \in B(\alpha x + (1-\alpha) y, d)$. It is now proven that F^{m+1} is concave, as it is a sum of a linear function and a concave function $F^{m,d}_{max}$. □

Lemma 37. *For each $m \in \{1,2,\ldots,N\}$ function F^m is piecewise linear on m intervals. Moreover, for some point f^*_m it holds that*

$$F^m(f) = m_m f + \begin{cases} F^{m-1}(f + d_{m-1}) & on\ [-1, f^*_m - d_{m-1}] \\ F^{m-1}(f^*_m) & on\ [f^*_m - d_{m-1}, f^*_m + d_{m-1}] \\ F^{m-1}(f - d_{m-1}) & on\ [f^*_m + d_{m-1}, 1] \end{cases} \tag{1.4}$$

Proof. The proof is conducted by induction over m. F^1 is a linear function, so it can be described by its values in $\{-1,1\}$. Assume that F^m can be described by at most $m+1$ points and is linear between these points. As F^m is concave, there exists a point

Figure 1.3: The method of constructing F^{m+1} from F^m



$f^*_m \in [-1,1]$ such that $F^m(f) \leq F^m(f^*_m)$ for every f. The maximum of F^m on an interval whose both ends are smaller than f^*_m is attained at its right end. Similarly, if both ends of the intervals are larger than f^*_m, the maximum is attained at its left end. Finally, if the interval contains f^*_m, the maximum is exactly at point x_m. These considerations prove the formula for F^{m+1}. Consequently, F^{m+1} is piecewise linear and it can be described by as many points as F^m plus 1. □

1.3.4.1. Pseudocode

The algorithm presented in this section constructs function F^N and finds its maximum. A set of pairs, called $funcDescription$, and a real variable $leftValue$ are used to represent F^{idx} for $idx \in \{1,2,...,N\}$. The structure has following interpretation:

1. $F^{idx}(-1) = leftValue$,
2. if $(v,p) \in funcDescription$ then $\frac{d}{dx}F^{idx}(x) = p$ for all x larger than v and smaller than the next value, v', in the structure.

For a given value v we define $\#v$ as $\min\{v' : (v',p) \in funcDescription \wedge v' > v\}$. By this definition $\frac{d}{dx}F^{idx}(x) = p$ on $(v,\#v)$ if $(v,p) \in funcDescription$.

Representation of F^0 is initialized to $F^0 \equiv 0$, namely

$$\begin{cases} leftValue = 0 \\ funcDescription = \{(-1,0),\ (1,-\infty)\} \end{cases} .$$

In each iteration of the main loop function F^{idx} is transformed into function F^{idx+1} as specified by equation (1.4). The transformation is achieved in three steps. Firstly, the maximum argument f^*_{idx} is found, all nodes on the left from f^*_{idx} are shifted to left, all nodes on the right from f^*_{idx} are shifted to the right, and a new node is added to represent the interval $[f^*_{idx}-d_{m-1}, f^*_{idx}+d_{m-1}]$. Secondly, value of $F^{idx}(-1)$ is computed and assigned to $leftValue$. Finally, the representation of F^{idx} is restricted to the interval $[-1,1]$ and linear function $m_m f$ is added.

`Input:`

- `non-decreasing table of positions,` $x \in \mathbb{R}^N$,
- `table of masses,` $m \in \mathbb{R}^N$.

`FLAT-DISTANCE (`$x \in \mathbb{R}^N$, $m \in \mathbb{R}^N$`):`
$leftValue \leftarrow 0$
$funcDescription \leftarrow \{(-1,0),\, (1,-\infty)\}$
`for` $idx \leftarrow 1$ `to` N `do`

$d \leftarrow x_{idx} - x_{idx-1}$
$funcLeft \leftarrow \{(v-d,p) : (v,p) \in funcDescription \wedge p > 0\}$
$funcRight \leftarrow \{(v+d,p) : (v,p) \in funcDescription \wedge p < 0\}$
$v_m \leftarrow \min\{v : (v,p) \in funcRight\}$
$funcDescription \leftarrow funcLeft \cup \{(v_m - 2d, 0)\} \cup funcRight$

$$leftValue \leftarrow leftValue + \sum_{\substack{(v,p)\in funcDescription \\ v<-1}} (\min(\#v,-1) - v)\, p$$

$(v_{min}, p_{min}) \leftarrow \max\{(v,p) : (v,p) \in funcDescription \wedge v \le -1\}$
$(v_{max}, p_{max}) \leftarrow \max\{(v,p) : (v,p) \in funcDescription \wedge v \in [-1,1]\}$
$funcDescription \leftarrow funcDescription \cap \{(v,p) : v \in (-1,1)\}$
$funcDescription \leftarrow funcDescription \cup \{(\max(v_{min}, -1), p_{min})\}$
$funcDescription \leftarrow funcDescription \cup \{(1,-\infty)\}$
$funcDescription \leftarrow \{(x, p + m_{idx}) : (x,p) \in funcDescription\}$

`return` $leftValue + \sum_{(v,p)\in funcDescription,\, p>0} (\#v - v) \cdot p$

Notice that the last instruction in the main loop, namely

$$funcDescription \leftarrow \{(x, p + m_{idx}) : (x,p) \in funcDescription\},$$

makes it inefficient to implement $funcDescription$ as a simple BST tree.

1.3.4.2. Flat-Distance in $\mathcal{O}(N^2)$

As mentioned before, the complexity of this algorithm depends on the implementation of $funcDescription$ data structure.

The simplest implementation of $funcDescription$ uses an array of pairs (v, p) sorted by v in ascending order and by p in the reverse order in the same time. This is possible as a consequence of Lemma 36.

The first block of instructions can be performed in $\Theta(\#funcDescription)$ by simply shifting all elements such that $p < 0$ to the right, and modifying v by iterating over all elements of $funcDescription$. The next block (computing of $leftValue$) can be computed with the same complexity, as

$$\min\{v' : (v', \_) \in funcDescription \wedge v' > v\}$$

is simply the next element after v in the ordered array. Finally, every instruction in the last block can be performed in $\Theta(\#funcDescription)$ by iterating over all its elements.

In each iteration of the main loop at most 1 element is added to $funcDescription$. Therefore, the computational complexity of the algorithm is $\mathcal{O}(N^2)$ while the memory complexity is $\mathcal{O}(N)$.

1.3.4.3. Flat-Distance in $\mathcal{O}(N \log N)$

The previous result can be improved to $\mathcal{O}(N \log N)$ by using balanced binary search trees data structure.

In this implementation $funcDescription$ is represented by global variable $p_{modifier}$ and a balanced binary search tree, T, of key-value pairs $(\Delta v, p)$ where p is the key. Let $\#p$ be the largest key in T smaller than p. The defined data structure $funcDescription$ specifies a function F^{idx} in the following sense:

1. $F^{idx}(-1) = leftValue$
2. if p is a key in T then $\frac{d}{dx}F^{idx}(x) = p + p_{modifier}$ for x such that

$$\sum_{\substack{(\Delta v', p') \in funcDescription \\ p' \geq p}} \Delta v' - 1 \leq x \leq \sum_{\substack{(\Delta v', p') \in funcDescription \\ p' \geq \#p}} \Delta v' - 1$$

Notice that obtaining a single element of $funcDescription$ (a pair (v, p) defining derivative in a given point) may take linear time.

The advantages of this structure can be easily seen when analyzing the first block of the code. The division of $funcDescription$ by the value of p (at first 0) can be achieved in $\mathcal{O}(\log N)$. Shifting all elements of those subsets can then be done in a constant time by modifying first elements of these sets. Adding the extra node also requires $\mathcal{O}(\log N)$ operations.

Setting $leftValue$ may require linear time, but all (apart from one) visited nodes in this process will be removed in the third block. Consequently the amortized cost of resetting $leftValue$ is $\mathcal{O}(N)$.

Removing nodes with the first coordinate $v < -1$ is obviously done in amortized $\mathcal{O}(N)$. Identifying nodes with the first coordinate $v > 1$ might seem problematic. It is, however, known that for the smallest p the respective v-value is equal to $1 + d$. Relevant nodes can be, therefore, removed in the reversed order (from right to left) $\mathcal{O}(N)$. Adding m_n to the second coordinate of each node is done simply by adding it to global variable $p_{modifier}$.

All iterations of the main loop require $\mathcal{O}(N \log N)$ operations. The memory complexity is also $\mathcal{O}(N \log N)$.

1.3.4.4. Performance of Flat-Distance implementations

Performance of the algorithm depends on the choice of *funcDescription* data structure. Theoretic bounds for computational complexity are, however, not sufficient to argue about performance of these two options. The first reason is that the each operation in $\mathcal{O}(N^2)$ algorithm is much faster than in $\mathcal{O}(N \log N)$ in terms of number of instructions. Secondly, hardware architectures provide solutions in which iterating over large tables is vastly accelerated. Finally, the algorithm does reach its theoretical bound only if many points concentrate on a small interval. A gap of size 2 between two points cleans *funcDescription* data structure completely. Numerical results presented in this section answer compare these two algorithms for different data input patterns. Performance was measured on a single core of AMD Athlon II X4 605e processor clocked at 2.3Ghz with 8GB of memory. The results are presented in Figures 1.4 and 1.5.

1.4. Comparison of metrics on $\mathfrak{M}^+(X)$

The following table presents a concise comparison of the distances defined in Section 1.1. For each metric basic properties, dual representation, compute complexity in the case of $X = \mathbb{R}$ and the distance between $2\delta_x$ and $3\delta_y$ are shown.

Metric	Example: $d(2\delta_x, 3\delta_y)$	Scale-invariance	Translation-invariance	Dual representation of $d(\mu, \nu)$	Compute complexity
Wasserstein	∞	YES	YES	The cost of optimal transference of distribution μ to ν, assuming that moving mass m by x requires mx energy.	$\mathcal{O}(N)$
Wasserstein normalized	$\min(2+3, (3-2)+\lvert x-y\rvert)$	weak	YES	Minimum of the sum of masses of μ and ν; and of the difference in masses between μ and ν plus the cost of transporting $\frac{\mu}{\lVert\nu\rVert}$ to $\frac{\nu}{\lVert\nu\rVert}$	$\mathcal{O}(N)$

Wasserstein centralized	$2\|x-y\|+\|y\|$	YES	NO	The difference in masses in point 0 in space added to the cost of transporting $\mu+(\|\nu\|-\|\mu\|)\,\delta_0$ to ν.	$\mathcal{O}(N)$
Flat	$1+$ $2\min(2,\|x-y\|)$	YES	YES	The cost of optimal transporting AND/OR generating AND/OR annihilating mass to form ν from μ.	$\mathcal{O}(N\log N)$
Radon	$2+3$	YES	YES	The cost of generating AND/OR annihilating mass to form ν from μ	$\mathcal{O}(N)$

Figure 1.4: Comparison of the performance of the two proposed algorithms for the flat distance between $0 \in \mathfrak{M}(\mathbb{R})$ and an N-point discrete measure with atoms randomly distributed over $[-1,1]$. The plot shows how the time of computation depends on N. For each input size 100 independent tests were executed to demonstrate how sensitive the algorithms are to input data distribution. Results of $\mathcal{O}(NlogN)$ algorithm are depicted as red dots, and results of $\mathcal{O}(N^2)$ algorithm as blue dots.

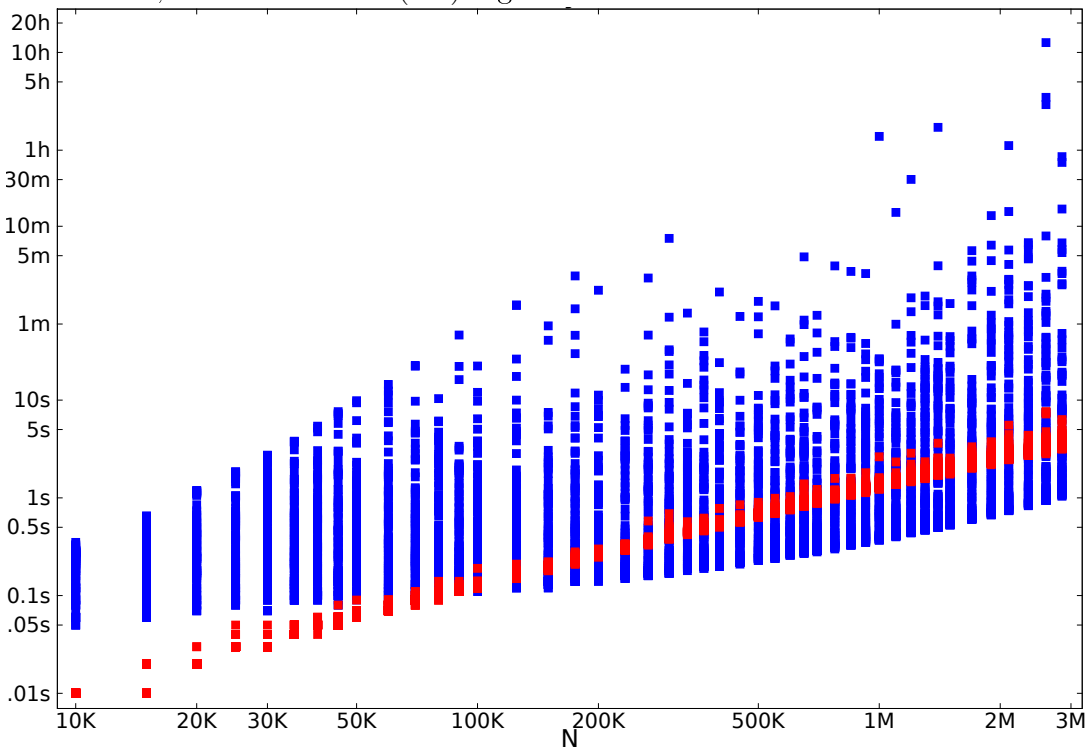
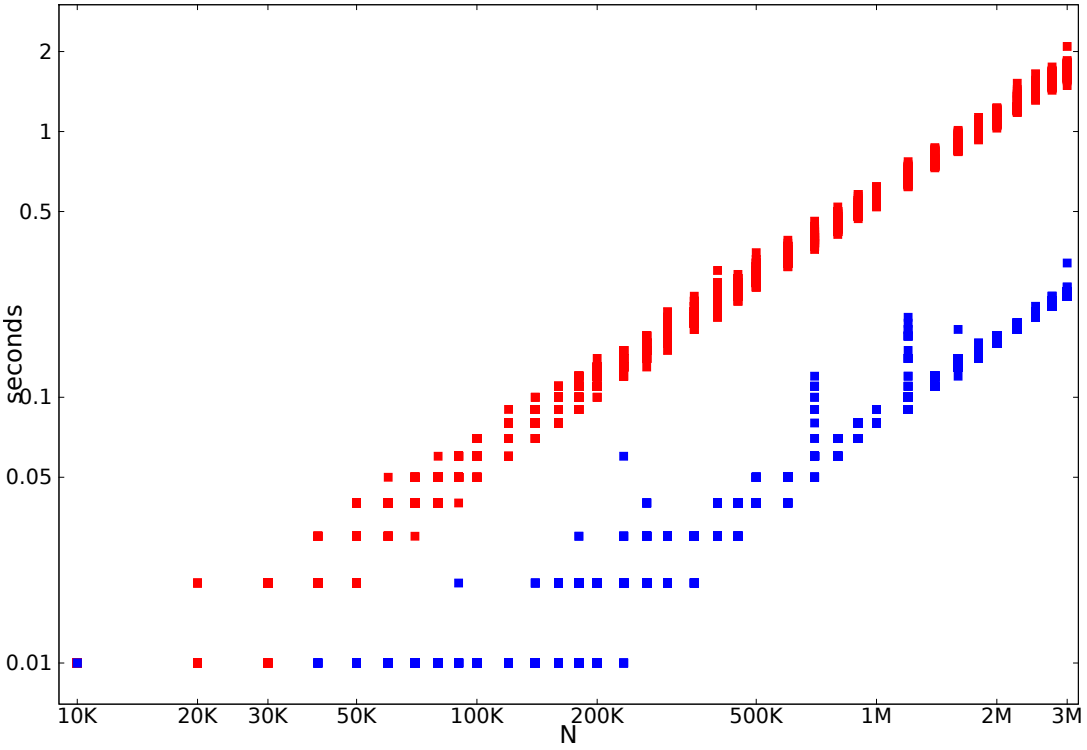


Figure 1.5: Comparison of the performance of the two proposed algorithms for the flat distance between $0 \in \mathfrak{M}(\mathbb{R})$ and a N-point discrete measure with atoms distributed over a large domain, i.e. distance between each two masses is larger than 2. In this case both algorithms are in fact linear, as the funcDescription structure has at most two elements. The plot demonstrates the overhead of using BST structures. Results of $\mathcal{O}(NlogN)$ algorithm are depicted as red dots, and results of $\mathcal{O}(N^2)$ algorithm as blue dots.



1.5. Approximation theory for Radon measures

In this section the following problem is considered: given a measure $\mu \in \mathfrak{M}^+[0,1]$ find its approximation $\mu^N \in \mathfrak{M}^+_{d,N}[0,1]$ supported on a N-element set which minimizes $\rho_F(\mu, \mu^N)$. Similar study has been presented in [68] and [63] for Wasserstein distance and the case of absolutely continuous measure μ. Theorem 23 allows to understand the approximation problems for Wasserstein distance as real-life questions about optimal choice of concentration points such as shops, warehouses and polling stations. Analogously, optimal flat distance approximation can be interpreted, following the lines of Corollary 26, as the optimal choice of concentration points of goods whose demand can be, at some additional cost, satisfied alternatively. Applications of this theory include the problem of locating parcel lockers and wireless access-points providing Internet services in metropolitan areas.

The results of this section are primarily motivated by the study of particle methods for solving partial differential equations (see Section 2.2). They prove to be useful for solving McKendrick-von Foerster equation numerically as they allow the following three improvements:

1. birth process can be implemented more efficiently,

2. number of particles, and hence the computational cost, can be reduced,

3. error of the scheme can be reduced by optimizing initial data approximation.

For details see Section 2.2. This section is divided into three subsections which cover general theory, approximations of discrete measures and approximations of absolutely continuous measures.

Remark 38. The choice of interval $X = [0, 1]$, often made in this section, is not arbitrary. Although, most of the results presented in this section can be generalized to the case of arbitrary interval $[a, b]$, it is not immediate. In many proofs (e.g. Theorem 44) it is necessary that the diameter of X does not exceed the Lipschitz constant of test functions in the definition of flat metric. Population dynamics equations considered in Section 2.2 can be rescaled to the interval $[0, 1]$, and hence results of this section can be applied without loss of generality.

1.5.1. Relation between 1-Wassersten and flat approximations

In many applications, where discrete measures are processed, it is desired to keep the number of atoms to the minimum while being sure that the introduced error is reasonable. This may require a compression step, where a measure consisting of N atoms is approximated by a M-point measure with $M < N$. In this section we investigate the bounds and asymptotics of the error induced by compressing discrete measures on $[0, 1]$.

Firstly, we define an equidistant N-point approximation. This method of approximation is efficient, easy to implement, and for some input measures it gives optimal results. Moreover it can be generalized to any totally bounded metric space. For certain applications, however, approximations of a better order can be constructed.

Lemma 39. *Let μ and ν be two non-negative measures on $[0, 1]$ with equal total variation, $\|\mu\| = \|\nu\|$, then $\rho_F(\mu, \nu) = W(\mu, \nu)$.*

Proof. Let $f_n \in Lip(1)$ be a sequence for which the following supremum is attained in the limit

$$W(\mu, \nu) = \sup \left\{ \left| \int_0^1 f d(\mu - \nu) \right| : f \in Lip(1) \right\},$$

then $\tilde{f}_n(x) = f_n(x) - f(0)$ is also a maximizing sequence, because

$$\int_0^1 f(0) d(\mu - \nu) = f(0) \cdot (\|\mu\| - \|\nu\|) = 0.$$

Conditions $\tilde{f} \in Lip(1)$ and $\tilde{f}(0) = 0$ imply that $|\tilde{f}(x)| \leq 1$ on $[0, 1]$ and hence

$$W(\mu, \nu) = \rho_F(\mu, \nu).$$

□

Definition 40. An equidistant N-point approximation, μ^N, of a non-negative Radon measure μ on $[a, b]$ is defined as

$$\mu^N = \sum_{i=0}^{N-2} \mu[a+(b-a)\frac{i}{N}, a+(b-a)\frac{i+1}{N})\delta_{a+(b-a)\frac{i+0.5}{N}} + \mu[a+(b-a)\frac{N-1}{N}, b]\delta_{b-(b-a)\frac{0.5}{N}}.$$

Theorem 41. *Let $\mu^N \in \mathfrak{M}_d[a,b]$ be an equidistant N-point approximation of $\mu \in \mathfrak{M}^+[a,b]$, then the following estimate holds*

$$\rho_F(\mu, \mu^N) \leq \frac{\mu[a,b]}{2N}$$

Proof. Measures μ and μ^N have equal total variation since

$$\left\|\mu^N\right\| = \sum_{i=0}^{N-2} \mu[a+(b-a)\frac{i}{N}, a+(b-a)\frac{i+1}{N}) + \mu[a+(b-a)\frac{N-1}{N}, b] = \mu[a,b] = \|\mu\|.$$

By Proposition 7 and Theorem 11

$$\rho_F(\mu, \mu^N) \leq \int_a^b |\mu[0,x] - \mu^N[0,x]|dx =$$

$$= \sum_{i=0}^{N-1} \int_{a+(b-a)\frac{i}{N}}^{a+(b-a)\frac{i+1}{N}} |\mu[a+(b-a)\frac{i}{N}, x] - \mu^N[a+(b-a)\frac{i}{N}, x]|dx =$$

$$= \sum_{i=0}^{N-1} \left(\int_{a+(b-a)\frac{i}{N}}^{a+(b-a)\frac{i+0.5}{N}} |\mu[a+(b-a)\frac{i}{N}, x]|dx + \int_{a+(b-a)\frac{i+0.5}{N}}^{a+(b-a)\frac{i+1}{N}} |\mu[x, a+(b-a)\frac{i+1}{N}]| \right) dx \leq$$

$$\leq \sum_{i=0}^{N-1} \left(\frac{b-a}{2N} \mu[a+(b-a)\frac{i}{N}, a+(b-a)\frac{i+1}{N})dx \right) =$$

$$\leq (b-a)\frac{\mu[a,b]}{2N}.$$

□

Corollary 42. *The subspace $\mathfrak{M}_d(\mathbb{R})$ is dense in $\mathfrak{M}(\mathbb{R})$.*

Proof. Let $\mu \in \mathfrak{M}^+(\mathbb{R})$ and let $\varepsilon > 0$. There exists $M \in \mathbb{N}$, such that

$$\mu[-M, M] \geq \mu(\mathbb{R}) - \frac{\varepsilon}{2}.$$

Let $N = 2M\mu(\mathbb{R})\varepsilon^{-1}$ and let μ^N be an N-point equidistant approximation of measure $\mu|_{[-M,M]}$. Then

$$\rho_F(\mu, \mu^N) \leq \rho_F(\mu, \mu|_{[-M,M]}) + \rho_F(\mu|_{[-M,M]}, \mu^N) \leq \varepsilon.$$

Since every finite Radon measure can be decomposed into a difference of two non-negative measures it completes the proof. □

Since we already know that any measure can be approximated arbitrarily well, we focus on the problem of optimal approximation. Lemma 39 and Theorem 44 guarantee that an optimal approximation in flat metric always exists and that it coincides with the optimal approximation in Wasserstein metric.

Definition 43. An optimal M-point discrete approximation, $\mu^M = \sum_{i=1}^M m_i \delta_{x_i}$, to a Radon measure μ is a discrete measure with M atoms minimizing

$$\rho_F(\mu, \mu^M)$$

Theorem 44. *Let $\mu \in \mathfrak{M}^+(\mathbb{R})$ and $supp\,\mu = X$. There exists an optimal M-point discrete approximation, μ^M, of μ in flat metric; it is supported on a subset of $conv\,X$, it is non-negative and moreover if $X \subseteq [0,1]$ then*

$$\|\mu\| = \left\|\mu^M\right\|.$$

Proof. The proof consists of five steps. In the first step we show existence of an optimal M-point approximation. In the next three steps we focus on the case discrete measures μ. In the second step we show that any M-point approximation of μ with atoms outside $conv\,X$. In the third step we show that any M-point approximation of μ whose total variation is different than the total variation of μ can be improved if μ is supported on a subset of $(0,1)$. In the fourth step we show that any M-point approximation of μ with negative masses can be improved. In the fourth step we generalize the results from steps 2-4 to the whole domain of non-negative Radon measures.

Step 1. Let $\{\mu_i^M\}_{i=1}^\infty$ be a sequence of M-point approximations such that

$$\rho_F(\mu, \mu_i^M) \to \inf\{\rho(\mu,\nu) : \nu \in \mathfrak{M}[0,1] \text{and } \nu \text{is an } M\text{-point discrete measure}\}$$

Each measure μ_i^M can be described by sequences $\{m_j^i\}_{j=1}^M$ and $\{x_j^i\}_{j=1}^M$ representing masses of Dirac deltas and their positions respectively. Namely,

$$\mu_i^M = \sum_{i=1}^M m_j^i \delta_{x_j^i}$$

By the compactness of $[0,1]^{2M} \subset \mathbb{R}^{2M}$ one can choose a subsequence $\{i_j\}_{j=1}^\infty \subseteq \{1,2,3,...\}$ such that

$$\forall_{i\in\{1,2,...,M\}} \quad m_i^{i_j} \to m_i \text{ and } x_i^{i_j} \to x_i.$$

By Lemma 19 the convergence of $\{m_i^{i_j}\}_{j=1}^\infty$ and $\{x_i^{i_j}\}_{j=1}^\infty$ implies that

$$\mu_{i_j}^M \to \sum_{i=1}^M m_i \delta_{x_i} = \mu^M \quad \text{in } \rho_F.$$

Consequently, μ^M is an optimal M-point approximation.

Step 2. In this step we assume that μ is a non-negative N-point discrete measure. We shall show that if μ^M is an M-point approximation of μ, not necessarily non-negative, then a better approximation, $\tilde{\mu}^M$ can be found provided that μ^M has atoms on $\mathbb{R} \setminus supp\,\mu$.

Firstly, we introduce some tools used in [40] for computing flat distance. Let

$$\mu - \mu^M = \sum_{i=1}^{M+N} m_i \delta_{x_i}$$

and $x_1 < x_2 < ... < x_{M+N}$. Define functions $\underline{F}^k : [-1,1] \to \mathbb{R}$ and $\overline{F}^k : [-1,1] \to \mathbb{R}$ by

$$\underline{F}^k(f) = \sup \left\{ \sum_{i=1}^{k} m_i f_i : \{f_i\}_{i=0}^{N+M} \subset [-1,1], f_k = f, \forall_{i \in \{1,..,k\}} |f_i - f_{i-1}| \leq |x_i - x_{i-1}| \right\},$$

$$\overline{F}^k(f) = \sup \left\{ \sum_{i=k}^{N+M} m_i f_i : \{f_i\}_{i=1}^{N+M} \subset [-1,1], f_k = f, \forall_{i \in \{k,...,N+M\}} |f_i - f_{i-1}| \leq |x_i - x_{i-1}| \right\}.$$

Obviously

$$\rho_F(\mu, \mu^M) = \sup_{f \in [-1,1]} \left| \underline{F}^{M+N}(f) \right| = \sup_{f \in [-1,1]} \left| \overline{F}^1(f) \right|. \tag{1.5}$$

Functions $\underline{F}^k$ and $\overline{F}^k$ are concave and piecewise linear (Lemma 36 and Lemma 37). From (1.3) it follows that

$$\underline{F}^{k+1}(f) = m_k f + \sup_{f_k \in [-1,1] \cap [f-(x_{k+1}-x_k), f+(x_{k+1}-x_k)]} \underline{F}^k(f_k). \tag{1.6}$$

We shall show that if μ^M has k atoms outside $supp\,\mu$ then a different approximation, $\tilde{\mu}^M_{k-1}$, at least as good as μ^M, consisting of at most $k-1$ atoms outside $supp\,\mu$ can be constructed.

Let $conv(supp\,\mu)) = [x_L, x_R]$. Suppose that μ^M has atoms outside $[x_L, x_R]$ and hence either $x_1 < x_L$ or $x_{N+M} > x_R$. Without loss of generality we can assume $x_{N+M} > x_R$. Let $\tilde{\mu}^M_{k-1} = \mu^M - m_{M+N}\delta_{x_{M+N}} + m_{M+N}\delta_{x_{M+N-1}}$. We have

$$\begin{aligned}
\rho(\mu, \mu^M) &= \sup_{f \in [-1,1]} \left| \underline{F}^{N+M}(f) \right| = \\
&= \sup_{f \in [-1,1]} \left| m_{N+M} f + \sup_{f_{N+M-1} \in [-1,1] \cap B(f, x_{N+M} - x_{N+M-1})} \underline{F}^{N+M-1}(f_{N+M-1}) \right| \geq \\
&\geq \sup_{f \in [-1,1]} \left| m_{N+M} f + \underline{F}^{N+M-1}(f) \right| = \rho(\mu, \tilde{\mu}^M_{k-1}).
\end{aligned}$$

The first non-trivial equality results from (1.6) and the estimate stems from the fact that $f \in B(f, x_{N+M} - x_{N+M-1})$. The last /equality follows from (1.5) and (1.6). Indeed, applying (1.6) to the case of measure $\mu - \mu^M - m_{M+N}\delta_{x_{M+N}} + m_{M+N}\delta_{x_{M+N-1}}$ instead of $\mu - \mu^M$ would result in formula

$$\underline{F}^{N+M} = m_{N+M} f + \underline{F}^{N+M-1}(f).$$

Step 3. In this step we assume that μ is a non-negative N-point discrete measure supported on a subset of $(0,1)$. We shall show that if μ^M is an M-point approximation

of μ, not necessarily non-negative, then $\tilde{\mu}_k^M = \mu^M + \left(\|\mu\| - \left\|\mu^M\right\|\right)\delta_{x_k}$ approximates μ at least as good as μ^M for any $k \in \{1, 2, ..., N+M\}$. To this end we introduce tools based on functions $\underline{F}^k$ and $\overline{F}^k$ for computing $\rho_F(\mu, \mu^M + \sum_{i=1}^{N+M} m_i \delta_{x_k})$.

Firstly, we will prove by induction the following statement: if $\mu - \mu^M$ is supported on a subset of $(0,1)$ then functions $\underline{F}^k$ and $\overline{F}^k$ are linear on $[-1+x_k, 1-x_k]$ and their derivatives at 0 are equal to $\sum_{i=1}^{k} m_i$ and $\sum_{i=k}^{N+M} m_i$ respectively. Let us focus on $\underline{F}^k$ as the case of $\overline{F}^k$ is fully analogical. $\underline{F}^1(f) = m_1 f$ is linear on $[-1+x_1, 1-x_1]$ and its derivative is equal to m_1. Suppose that F^k is linear on $[-1+x_k, 1-x_k]$ and its derivative at 0 equals $\sum_{i=1}^{k} m_i$. From (1.6) we have

$$\underline{F}^{k+1}(f) = m_k f + \sup_{f_k \in [-1,1] \cap B(f, x_{k+1} - x_k)} \underline{F}^k(f_k).$$

Since $\underline{F}^k(f_k)$ is linear on $[-1+x_k, 1-x_k]$ it follows that for all $x \in [-1+x_{k+1}, 1-x_{k+1}]$ we have either

$$\sup_{f_k \in [-1,1] \cap B(f, x_{k+1} - x_k)} \underline{F}^k(f_k) = \underline{F}^k(f + (x_{k+1} - x_k))$$

or

$$\sup_{f_k \in [-1,1] \cap B(f, x_{k+1} - x_k)} \underline{F}^k(f_k) = \underline{F}^k(f - (x_{k+1} - x_k)).$$

Hence, $\underline{F}^{k+1}(f)$ is linear on $[-1+x_{k+1}, 1-x_{k+1}]$ and $\frac{d}{df}\underline{F}^{k+1}(0) = m_k + \frac{d}{df}\underline{F}^k(0)$. This proves the inductive step.

For $k \in \{1, 2, ..., N+M\}$ we define

$$\begin{aligned} G(x_k, f) &= \sup_{y \in [-1,1] \cap B(f, x_k - x_{k-1})} \underline{F}^{k-1}(y) + m_k f + \sup_{y \in [-1,1] \cap B(f, x_{k+1} - x_k)} \overline{F}^{k+1}(y) = \\ &= \underline{F}^k(y) - m_k f + \overline{F}^k(y). \end{aligned} \tag{1.7}$$

Notice that for any $x \in [0,1]$ and $m \in \mathbb{R}$ we have

$$\rho(\mu, \mu^M + m\delta_{x_k}) = \sup_{f \in [-1,1]} \left| G(x_k, f) - mf \right|.$$

Since both functions $\underline{F}^k$ and $\overline{F}^k$ are concave on $[-1,1]$ and linear on $(-\varepsilon, \varepsilon)$ we conclude that so is $G(x, \cdot)$. Finally,

$$\frac{d}{df} G(x_k, 0) = \frac{d}{df}\overline{F}^k(0) + \frac{d}{df}\underline{F}^k(0) - m_k = \sum_{i=1}^{N+M} m_i. \tag{1.8}$$

From $\|\mu\| \neq \left\|\mu^M\right\|$ we have

$$\frac{d}{df} G(x_k, 0) = \sum_{i=1}^{N+M} m_i \neq 0.$$

Observe that

$$\|\mu\| = \left\| \mu^M + \left(\sum_{i=1}^{N+M} m_i \right) \delta_{x_k} \right\| = \left\| \tilde{\mu}_k^M \right\|.$$

Using the fact that $G(x_k,\cdot)-\left(\sum_{i=1}^{N+M} m_i\right) f$ attains its maximum at $f=0$ and consequently that $G(x_k,\cdot)$ attains its maximum outside $[-\varepsilon,\varepsilon]$ we obtain

$$\rho_F(\mu,\tilde{\mu}_k^M)=\rho_F\left(\mu,\mu^M+\left(\sum_{i=1}^{N+M} m_i\right)\delta_{x_k}\right)=\sup_{f\in[-1,1]}\left[G(x_k,f)-\left(\sum_{i=1}^{N+M} m_i\right)f\right]=$$

$$=G(x_k,0)<\sup_{f\in[-1,1]} G(x_k,f)=\rho(\mu,\mu^M).$$

This completes the proof of the third step.

Step 4. In this step we assume that μ is a non-negative N-point discrete measure with no atoms in $\{0,1\}$. We shall show that if μ^M is an M-point approximation of μ, not necessarily non-negative, satisfying $\|\mu\|=\left\|\mu^M\right\|$ then a better, non-negative approximation, $\tilde{\mu}^M$, can be found and also $\|\mu\|=\left\|\tilde{\mu}^M\right\|$.

Let

$$\begin{aligned}\mu &= \sum_{i=1}^{N} m_i\delta_{x_i},\\ \left(\mu^M\right)^- &= -\sum_{i=N+1}^{K} m_i\delta_{x_i},\\ \left(\mu^M\right)^+ &= \sum_{i=K+1}^{N+M} m_i\delta_{x_i}.\end{aligned}$$

From Lemma 39

$$\rho_F(\mu,\mu^M)=W\left(\mu-\left(\mu^M\right)^-,\left(\mu^M\right)^+\right).$$

Let $\gamma^*\in\Gamma\left(\mu-\left(\mu^M\right)^-,\left(\mu^M\right)^+\right)$ be the optimal transference plan. For $k\in\{N+1,...,K\}$ let

$$\nu_k=\frac{m_k}{\|\gamma^*(\{x_k\},\cdot)\|}\gamma^*(\{x_k\},\cdot).$$

Define

$$\tilde{\mu}^M=\left(\mu^M\right)^+-\sum_{i=N+1}^{K}\nu_k,$$

$$\tilde{\gamma}(E,\cdot)=\gamma^*(E,\cdot)-\sum_{\substack{i\in\{N+1,...,K\}\\ x_i\in E}}\nu_k.$$

Newly defined $\tilde{\mu}^M$ is an M-point approximation since

$$supp\nu_k=supp\gamma^*(\{x_k\},\cdot)\subset supp\gamma^*([0,1],\cdot)=supp\mu^M.$$

and it is non-negative because $m_k\le\|\gamma^*(\{x_k\},\cdot)\|$. Also $\|\mu\|=\left\|\tilde{\mu}^M\right\|$. Since $\gamma^*-\tilde{\gamma}\neq 0$ and $(\gamma^*-\tilde{\gamma})\in\Gamma(\mu,\tilde{\mu}^M)$ we obtain

$$\rho_F(\mu,\tilde{\mu}^M)=W(\mu,\tilde{\mu}^M)\le\int_{[0,1]^2}|x-y|d\tilde{\gamma}<\int_{[0,1]^2}|x-y|d\gamma^*=W\left(\mu-\left(\mu^M\right)^-,\left(\mu^M\right)^+\right),$$

which completes the proof.

Step 5. In the last step of the proof we shall generalize the reasoning from Step 2 and Step 3 from the case of μ being a non-negative discrete measure supported on a subset of $[\varepsilon, 1-\varepsilon]$ to the case of an arbitrary measure $\mu \in \mathfrak{M}^{\geq}[0,1]$. By Theorem 41 for every $\varepsilon > 0$ there exists a discrete measure μ_ε with no atoms in $\{0,1\}$ such that $\rho_F(\mu, \mu_\varepsilon) \leq \varepsilon \|\mu\|$ and $\|\mu\| = \|\mu_\varepsilon\|$. Denote the optimal M-point approximation of μ_ε by μ_ε^M. We have

$$\rho_F(\mu_\varepsilon, \mu_\varepsilon^M) \leq \rho_F(\mu_\varepsilon, \mu^M) \leq \rho_F(\mu_\varepsilon, \mu) + \rho_F(\mu, \mu^M) \leq \varepsilon \|\mu\| + \rho_F(\mu, \mu^M).$$

There exists a sequence $\varepsilon_n \to 0$ for which $\mu_{\varepsilon_n}^M$ is convergent. Let $\mu_{\varepsilon_n}^M \to \tilde{\mu}^M$. Finally from

$$\rho_F(\mu, \mu_\varepsilon^M) \leq \rho_F(\mu, \mu_\varepsilon) + \rho_F(\mu_\varepsilon, \mu_\varepsilon^M) \leq 2\varepsilon \|\mu\| + \rho_F(\mu, \mu^M)$$

we obtain

$$\rho_F(\mu, \tilde{\mu}^M) \leq \rho_F(\mu, \mu_\varepsilon^M) + \rho_F(\mu_\varepsilon^M, \tilde{\mu}^M) \to \rho_F(\mu, \mu^M),$$

hence $\tilde{\mu}^M$ is the optimal N-point approximation of μ. □

Theorem 44 provides a strong tool for dealing with optimal approximation problems. It is used in the Lemma below for constructing an optimal 1-point approximation of an arbitrary measure.

Definition 45. Let μ be a non-negative measure on $[a,b] \subseteq [0,1]$. We define the central point of measure μ as

$$x^*_{[a,b]} = \sup\{x \in [a,b] : \mu[a,x) \leqslant \mu(x,b]\}.$$

Lemma 46. *Let μ be a non-negative measure on $[a,b] \subseteq [0,1]$, then $\nu_{x^*_{[a,b]}} = \|\mu\| \, \delta_{x^*_{[a,b]}}$ is an optimal 1-point approximation of μ in flat metric and*

$$\rho_F(\mu, \nu_{x^*_{[a,b]}}) = \int_0^{x^*_{[0,1]}} \mu[0,x]dx + \int_{x^*_{[0,1]}}^1 \mu[x,1]dx.$$

Proof. By Theorem 44 the mass of the optimal approximation of μ equals $\|\mu\|$. Let $\nu_x = \|\mu\| \, \delta_x$ be the optimal approximation of μ, then by Lemma 39 and Theorem 11

$$\rho_F(\mu, \nu_x) = W(\mu, \nu_x) = \int_0^1 |\mu[0,\tau] - \nu_x[0,\tau]|d\tau = \int_0^x \mu[0,\tau]d\tau + \int_x^1 \mu[\tau,1]d\tau.$$

Suppose to the contrary without loss of generality that $x > x^*_{[0,1]}$. Consequently,

$$\rho_F(\mu, \nu_x) - \rho_F(\mu, \nu) = \int_{x^*_{[0,1]}}^x \mu[0,\tau] - \mu[\tau,1]d\tau. \tag{1.9}$$

By the definition of $x^*_{[a,b]}$ the integrant in (1.9) is a strictly positive function, which contradicts optimality of ν_x. □

Proposition 47. *The estimate $\rho_F(\mu, \nu_{x^*_{[0,1]}}) \leqslant \frac{1}{2}\mu[0,1]$ holds. Equality is satisfied for $\mu = \delta_0 + \delta_1$.*

Proof. The estimate follows immediately from Theorem 41. Consider $\mu = \delta_0 + \delta_1$, then $x^*_{[0,1]} = 1$ and

$$\rho_F(\mu, v_{x^*_{[0,1]}}) = W(\delta_0 + \delta_1, 2\delta_1) = \frac{1}{2}\mu[0,1] = 1.$$

□

The previous result may seem a little disappointing because in the worst case the error of a 1-point optimal approximation is exactly equal to the error of a 1-point equidistant approximation. In the next section we focus on the problem of finding an $N-1$-point approximation to a given N-point discrete measure on $[0,1]$. It turns out that there exists a linear algorithm for finding optimal approximation in this case and that the upper bound for the error is of order N^{-2}.

1.5.2. Reduction of the number of atoms in a discrete measure

Lemma 48. *Let $\mu = \sum_{i=1}^N m_i\delta_{x_i}$ be a non-negative measure on $[0,1]$, then there exists an optimal M-point approximation supported on a subset of $\{x_i\}_{i=1}^N$.*

Proof. Let $\mu^M = \sum_{i=1}^M n_i\delta_{y_i}$ be an optimal M-point approximation that is not supported on a subset of $\{x_i\}_{i=1}^N$. By Theorem 44 μ^M is supported on a subset of $[x_1, x_N]$ and $\sum_{i=1}^M n_i = \sum_{i=1}^N m_i$. Suppose that for some indices a, b, c it holds

$$x_a < y_b < y_{b+1} < ... < y_{b+c} < x_{a+1}.$$

By Theorem 39 and Theorem 11

$$\rho_F(\mu, \mu^M) = \phi + \psi(\{y_i\}_{i=b}^{b+c}),$$

where

$$\phi = \int_{[0,x_a]\cup[x_{a+1},1]} |\mu[0,x] - \mu^M[0,x]|dx,$$
$$\psi(\{y_i\}_{i=b}^{b+c}) = \int_{x_a}^{x_{a+1}} |\mu[0,x] - \mu^M[0,x]|dx.$$

Because μ^M is a discrete measure we have

$$\psi(\{y_i\}_{i=b}^{b+c}) = \int_{x_a}^{y_b} |\mu[0,x_a] - \mu^M[0,x_a]|dx + \int_{y_b}^{y_{b+1}} |\mu[0,x_a] - \mu^M[0,y_b]|dx + ...+$$
$$+ \int_{y_{b+c}}^{x_{a+1}} |\mu[0,x_a] - \mu^M[0,y_{b+c}]|dx$$

or simply

$$\psi(\{y_i\}_{i=b}^{b+c}) = (y_b - x_a)|\mu[0,x_a] - \mu^M[0,x_a]| + ... + (x_{a+1} - y_{b+c})|\mu[0,x_a] - \mu^M[0,y_{b+c}]|,$$

which implies

$$\psi(\{y_i\}_{i=b}^{b+c}) \geq (x_{a+1} - x_a) \cdot \min \left(\left\{ |\sum_{i=1}^{a} m_i - \sum_{i=1}^{d} n_i| \right\}_{d=b-1}^{b+c} \right). \tag{1.10}$$

Suppose the minimum is attained at index $d = D$. Define measure $\tilde{\mu}^M$ as

$$\tilde{\mu}^M = \sum_{i=1}^{b-1} n_i \delta_{y_i} + \left(\sum_{i=b}^{D} n_i \right) \delta_{x_a} + \left(\sum_{i=D+1}^{b+c} n_i \right) \delta_{x_{a+1}} + \sum_{i=b+c+1}^{M} n_i \delta_{y_i}.$$

Notice that $\tilde{\mu}^M$ is concentrated on a set of cardinality at most M. The error of approximation is given by

$$\rho_F(\mu, \tilde{\mu}^M) = \phi + \int_{x_a}^{x_{a+1}} |\sum_{i=1}^{a} m_i - \sum_{i=1}^{D} n_i| dx,$$

which by inequality (1.10) can be estimated from above by $\phi + \psi(\{y_i\}_{i=b}^{b+c})$ and consequently

$$\rho_F(\mu, \mu^M) \geq \rho_F(\mu, \tilde{\mu}^M).$$

It proves that there exists an optimal M-point approximation to μ that is supported on a set with no points in $\bigcup_{i=1}^{N-1}(x_i, x_{i+1})$. □

On the basis of Theorem 44 and Lemma 48 a brute-force algorithm for finding optimal $N-1$-point approximation can be built. The idea is to compute minimal error in flat metric for each possible support.

Definition 49. Given a discrete, non-negative measure $\mu = \sum_{i=1}^{N} m_i \delta_{x_i}$ with N atoms on $[0, 1]$ we define $N-1$-point approximation algorithm as follows:

1. For each $j = 1, 2, ..., N$ consider the set $\hat{x}_j = \{x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_N\}$ and define

$$\mu_j^L = \sum_{i=1}^{j-1} m_i \delta_{x_i} + m_j \delta_{x_{j-1}} + \sum_{i=j+1}^{N} m_i \delta_{x_i},$$
$$\mu_j^R = \sum_{i=1}^{j-1} m_i \delta_{x_i} + m_j \delta_{x_{j+1}} + \sum_{i=j+1}^{N} m_i \delta_{x_i},$$

2. For each side $s \in \{L, R\}$ compute

$$\rho_F(\mu, \mu_i^s).$$

3. Return the measure μ_i^s which accounts for the lowest error $\rho_F(\mu, \mu_i^s)$.

Proposition 50. *The $N-1$-point approximation algorithm has linear computational complexity.*

Proof. By Lemma 39 and Theorem 11

$$\rho_F(\mu, \mu_i^L) = W(\mu, \mu_i^L) = m_i(x_i - x_{i-1}),$$
$$\rho_F(\mu, \mu_i^R) = W(\mu, \mu_i^R) = m_i(x_{i+1} - x_i).$$

Consequently, the value of $\rho_F(\mu, \mu_i^s)$ can be computed in constant time for any $i \in \{1, ..., N\}$ and $s \in \{L, R\}$ and thus the algorithm requires $\mathcal{O}(n)$ operations. □

Theorem 51. *For a non-negative N-point measure μ the $N-1$-point approximation algorithm returns an optimal $N-1$-point approximation, μ^{N-1}, and the following estimate holds*

$$\rho_F(\mu, \mu_{N-1}) \leq \frac{2\,\|\mu\|}{N^2}. \tag{1.11}$$

Proof. Let $\mu = \sum_{i=1}^N m_i \delta_{x_i}$ be a non-negative measure on $[0, 1]$ and let

$$\mu^{N-1} = \sum_{i=1}^{k-1} n_i \delta_{x_i} + \sum_{k+1}^{N} n_i \delta_{x_i}$$

be its $N-1$-point optimal approximation. Denote $\Delta_i = n_i - m_i$ for all $i \neq k$ and $\Delta_k = -m_k$. It's easy to show that $\Delta_i \geq 0$ for $i \neq k$. Consequently from Lemma 44 and Theorem 11 we derive a formula for ρ_F and from non-negativity of Δ_k we can deduce the sign of $\sum_{j=1}^i \Delta_i$ and therefore omit the absolute value:

$$\rho_F(\mu, \mu^{N-1}) = \sum_{i=1}^{k-1} \left((x_{i+1} - x_i) \sum_{j=1}^{i} \Delta_j \right) + \sum_{i=k}^{N} \left(-(x_{i+1} - x_i) \sum_{j=1}^{i} \Delta_j \right).$$

By changing the order of summation we obtain

$$\rho_F(\mu, \mu^{N-1}) = \sum_{i=1}^{N} \Delta_i |x_k - x_i|.$$

This value is minimal when $\Delta_i = 0$ hold for all i except for $i = k-1$ or $i = k+1$. The $N-1$-point approximation algorithm indeed considers all measures concentrated on $\{x_i\}_{i=1}^N$ that satisfy this condition and $\left\|\mu^{N-1}\right\| = \left\|\mu^N\right\|$.

To prove the estimate 1.11 let us denote $d_i = x_{i+1} - x_i$. The distance between μ and μ_{N-1} is equal to the cost of moving some mass m_i from one of the neighboring nodes

$$\rho_F(\mu, \mu_{N-1}) = W(\mu, \mu_{N-1}) = \min \left\{ m_1 d_1, \min_{i \in \{2,\dots,N-1\}} m_i \cdot \min(d_i, d_{i-1}),\, m_N d_{N-1} \right\}$$

and consequently

$$\rho_F(\mu, \mu_{N-1}) \leqslant \min \left\{ \min_{i \in \{1,\dots,N-1\}} m_i \cdot d_i,\, m_N d_{N-1} \right\}.$$

Using Schwarz inequality for sequences $\left\{\sqrt{\frac{m_i}{\|\mu\|}}\right\}_{i=1}^{N}$ and $\left\{\sqrt{d_{\min(i,N-1)}}\right\}_{i=1}^{N}$ we obtain

$$\left(\sum_{i=1}^{N-1}\sqrt{\frac{m_i d_i}{\|\mu\|}}+\sqrt{\frac{m_N d_{N-1}}{\|\mu\|}}\right)^2 \leq \left(\sum_{i=1}^{N}\frac{m_i}{\|\mu\|}\right)\left(\sum_{i=1}^{N-1} d_i + d_{N-1}\right) \leq 2,$$

since

$$\sum_{i=1}^{N-1} m_i \leqslant \|\mu\| \quad \text{and} \quad \sum_{i=1}^{N-1} d_i \leqslant 1$$

and

$$\sum_{i=1}^{N-1}\sqrt{m_i d_i}+\sqrt{m_N d_{N-1}} \leq \sqrt{2\,\|\mu\|}.$$

It implies that there exists i such that

$$\sqrt{m_i d_i} \leq \frac{\sqrt{2\,\|\mu\|}}{N} \quad \text{or} \quad \sqrt{m_N d_{N-1}} \leq \frac{\sqrt{2\,\|\mu\|}}{N}$$

and consequently

$$m_i d_i \leq \frac{2\,\|\mu\|}{N^2} \quad \text{or} \quad m_N d_{N-1} \leq \frac{2\,\|\mu\|}{N^2}.$$

□

Theorem 51 guarantees that removing a single atom from an N-point measure and readjusting masses does not induce a large error. Consequently removing a fixed number k of atoms induces an error of order N^{-2} as well. Obviously, if k is proportional to N then the estimate from the Theorem 51 only guarantees the error of order N^{-1}. The examples below show that in such case ($k \sim N$) no better estimate of the error of optimal approximation can be found and that applying $N-1$-point approximation algorithm iteratively k-times does not lead to good results.

Remark 52. Algorithm for finding M-point approximation of N-point discrete measure by removing optimally one mass at a time (greedy algorithm) is suboptimal. Indeed, let us consider measure $\mu = \sum_{n=1}^{N} \frac{1}{n}\delta_{1/n}$ then a 1-point approximation constructed by removing one mass at a time is given by $\mu^1 = \left(\sum_{n=1}^{N}\frac{1}{n}\right)\delta_1$ while the optimal 1-point approximation tends to $\left(\sum_{n=1}^{N}\frac{1}{n}\right)\delta_0$ with $N \to \infty$. The error of the greedy algorithm therefore can be higher then the error of the equidistant approximation.

Theorem 53. *For every $M < N$ there exists an N-point discrete measure on $[0,1]$ whose optimal M-point approximation yields error equal to*

$$\|\mu\|\,\frac{N-M}{N(N-1)}.$$

Proof. Let

$$\mu^N = C \sum_{i=0}^{N-1} \frac{1}{N} \delta_{i/N-1},$$

where $C > 0$ is any constant, and μ^M be its optimal M-point approximation. By Lemma 44, Lemma 39 and Theorem 23 the error of the approximation equals

$$\rho_F(\mu^N, \mu^M) = W(\mu^N, \mu^M) = \inf_{\gamma \in \Gamma(\mu^N, \mu^M)} \left\{ \int_{[0,1]^2} |x-y| d\gamma \right\}.$$

By Lemma 48 the set $supp\, \mu^N \backslash supp\, \mu^M$ consists of exactly $N-M$ points: $x_{j_1}, x_{j_2}, ..., x_{j_{N-M}}$. Obviously for every $k \in \{1, 2, ..., N-M\}$ and $x \in supp\, \mu^M$ the distance $|x - x_{j_k}| \geq \frac{1}{N-1}$. On the other hand for every transference plan $\gamma \in \Gamma(\mu^N, \mu^M)$ the mass transported from the point x_k equals

$$\gamma(\{x_{j_k}\}, [0,1]) = \frac{C}{N}.$$

As there are $N-M$ Dirac deltas with masses $\frac{C}{N}$ each that have to be shifted by the distance at least $\frac{1}{N-1}$ we can conclude that

$$\rho_F(\mu^N, \mu^M) \geq \inf_{\gamma \in \Gamma(\mu^N, \mu^M)} \left\{ \sum_{i=1}^{M} \sum_{k=1}^{N-M} |x_{j_k} - x_i| \gamma\left(\{x_{j_k}\} \times \{x_i\}\right) \right\} \geq$$

$$\geq \frac{C}{N} \sum_{k=1}^{N-M} |x_{j_k} - x_i| \geq \frac{C}{N} \sum_{k=1}^{N-M} \frac{1}{N-1} \geq C \frac{N-M}{N(N-1)}$$

□

Remark 54. If M be proportional to N then the error of an optimal M-point approximation to a N-point measure can be of the same order as the error of an M-point equidistant approximation.

1.5.3. Approximation of absolutely continuous measures on $[0,1]$

For any absolutely continuous measure an equidistant M-point approximation can be built. It induces an error of order M^{-1} and it is optimal in the case of a constant function. Nonetheless, for many other applications (such as multi-hump functions) this approximation can be improved by a large factor. In this section we investigate methods of improving discrete approximations of absolutely continuous measures.

In the beginning of this section we shall recall some observations from [68], which constitute an excellent tool for improving the error of approximation. Farther, we introduce two new methods and compare the results against the algorithm investigated in [68].

Definition 55. Let $\mathscr{M} : L^1(X) \to \mathcal{M}(X)$ be an inclusion map given by

$$\mathscr{M}(f)(E) = \int_E f d\mathcal{L},$$

where $\mathcal{L}$ is the Lebesgue measure.

The flat distance between a Radon measure μ on a normed space X and a function $f \in L^1(X)$ is defined as $\rho_F(\mu, f) = \rho_F(\mu, \mathscr{M}(f))$. Similarly for $f, g \in L^1(x)$ we define $\rho_F(f, g) = \rho_F(\mathscr{M}(f), \mathscr{M}(g))$.

Let us consider a discrete M-point approximation μ^M of a positive continuous function $f \in C[0,1]$. The domain $[0,1]$ can be divided into M sets corresponding to the areas to which each of Dirac mass of μ^M is transported. It turns out that if μ^M is the optimal approximation of f then this division is a partition of $[0,1]$. The following definition and Proposition 57 provide precise formulation of this intuition.

Definition 56. Let $f \in C[0,1]$ and let $\mu^M = \sum_{i=1}^M m_i \delta_{x_i}$ be any M-point approximation such that $\left\|\mu^M\right\| = \|\mathscr{M}(f)\|$. Let $\gamma^* \in \Gamma(\mu^M, \mathscr{M}(f))$ be the optimal transference plan (see Theorem 24). We define the *transport domain division* of interval $[0,1]$ as a sequence of sets X_i such that

$$X_i = supp\left(\gamma^*(\{x_i^*\}, \cdot)\right)$$

for $i = 1, 2, ..., M$.

Proposition 57. *Let $\{X_i\}_{i=1}^M$ be the transport domain division of interval $[0,1]$ for $f \in C[0,1]$ and its optimal M-point approximation $\mu^M = \sum_{i=1}^M m_i \delta_{x_i^*}$. Suppose that the sequence $\{x_i^*\}_{i=1}^M$ is increasing. The following statements hold*

1. *$x_i^* \in X_i$ for every $i = 1, 2, ..., M$,*
2. *for every $i = 1, 2, ..., M$ it holds $X_i = [a_i, a_{i+1}]$ with $a_1 = 0$ and $a_{M+1} = 1$,*
3. *$a_{i+1} - x_i^* = x_{i+1}^* - a_{i+1}$ for every $i = 1, 2, ..., M-1$.*

Proof. Let γ^* be the optimal transference plan of μ^M to $\mathscr{M}(f)$. Since

$$supp(\gamma^*([0,1], \cdot)) = supp(\mathscr{M}(f)) = [0,1]$$

and

$$supp(\gamma^*([0,1], \cdot)) = \bigcup_{i=1}^M supp(\gamma^*(\{x_i^*\}, \cdot)) = \bigcup_{i=1}^M X_i$$

we conclude that

$$\bigcup_{i=1}^M X_i = [0,1]. \tag{1.12}$$

Next, we prove that condition $|x - x_j^*| < |x - x_k^*|$ implies $x \notin X_k$. Suppose to the contrary that $x \in X_k$ and there exists a neighborhood $N_x \ni x$ such that for every $y \in N_x$ we have

$$|y - x_j^*| < |y - x_k^*|. \tag{1.13}$$

Let us define an alternative approximation, $\tilde{\mu}^M$, of f by

$$\tilde{\mu}^M = \sum_{\substack{i \in \{1,2,...,M\} \\ i \neq j,\ i \neq k}} m_i \delta_{x_i^*} + (m_j + \Delta)\, \delta_{x_j^*} + (m_k - \Delta)\, \delta_{x_k^*},$$

$$\Delta = \gamma^*(\{x_k\}, N_x),$$

and a transference plan, $\tilde{\gamma}$, by

$$\tilde{\gamma}(A,B) = \begin{cases} \gamma^*(A,B) & \text{if } x_j, x_k \notin A \text{ or } x_j, x_k \in A \\ \gamma^*(A,B) - \gamma^*\left(\{x_k\}, B \cap N_x\right) & \text{if } x_j \notin A,\ x_k \in A \\ \gamma^*(A,B) + \gamma^*\left(\{x_k\}, B \cap N_x\right) & \text{if } x_k \notin A,\ x_j \in A. \end{cases}$$

Notice that $\tilde{\gamma} \in \Gamma(\tilde{\mu}^M, \mathscr{M}(f))$. Using 1.13 we obtain

$$\rho_F(\mu^M, f) - \rho_F(\tilde{\mu}^M, f) \geq \int_{[0,1]^2} |x-y| d\left(\gamma^* - \tilde{\gamma}\right) =$$

$$= \int_{N_x} |x_j - y| d\left(\gamma^*(\{x_j\}, \cdot) - \tilde{\gamma}(\{x_j\}, \cdot)\right) + \int_{N_x} |x_k - y| d\left(\gamma^*(\{x_k\}, \cdot) - \tilde{\gamma}(\{x_k\}, \cdot)\right) =$$

$$> \int_{N_x} |x_j - y| d\left(\gamma^*(\{x_j, x_k\}, \cdot) - \tilde{\gamma}(\{x_j, x_k\}, \cdot)\right) = 0,$$

which contradicts optimality of μ^M.

Statement (1) easily follows from the fact that condition $0 = |x_j^* - x_j^*| < |x_j^* - x_k^*|$ implies $x_j \notin X_k$, which is holds for every $k \neq j$.

To prove statement (2) suppose X_j is not an interval; equivalently there exist $x_j, x_j' \in X_j$ and $x_k \in (x_j, x_j')$ such that $x_k \in X_k$. These assumptions imply following inequalities: $|x_j - x_j^*| < |x_j - x_k^*|$, $|x_j' - x_j^*| < |x_j' - x_k^*|$ and $|x_k - x_j^*| > |x_k - x_k^*|$. From the first two inequalities we have $x_k^* \notin [x_j, x_j']$, which contradicts the third inequality. Since $\bigcup_{i=1}^M X_i = [0,1]$ and $x_i^* \in X_i$, then indeed $a_1 = 0$ and $a_{M+1} = 1$.

Statement (3) follows from the fact that $a_{i+1} \in X_i$ and $a_{i+1} \in X_{i+1}$ hence neither $|a_{i+1} - x_i^*| < |a_{i+1} - x_{i+1}^*|$ nor $|a_{i+1} - x_i^*| > |a_{i+1} - x_{i+1}^*|$ holds. □

Definition 58. Let $\{X_i\}_{i=1}^M$ be the transport domain division of interval $[0,1]$ corresponding to an M-point discrete measure μ^M. If $\{X_i\}_{i=1}^M = \{[a_i, a_{i+1}]\}_{i=1}^M$ and

$$0 = a_1 < a_2 < ... < a_M < a_{M+1} = 1,$$

then the sequence $\{a_1, a_2, ..., a_{M+1}\}$ is called a transport partition of interval $[0,1]$ corresponding to μ^M.

The following fact follows immediately from Lemma 46.

Corollary 59. *For a given sequence $\{a_i\}_{i=1}^{M+1} \subset [0,1]$ and a positive continuous function f the M-point approximation of f which is optimal in the class of measures whose transport partition coincides with $\{a_i\}_{i=1}^{M+1}$ is given by*

$$\mu^M = \sum_{i=1}^{M} \int_{[a_i, a_{i+1}]} f(x)dx \cdot \delta_{x^*_{[a_i, a_{i+1}]}},$$

where $x^*_{[a,b]}$ is the central point of measure $\mathscr{M}(f)|_{[a,b]}$, see Definition 45.

Proof. Let $\mu^M = \sum_{i=1}^M m_i \delta_{x_i}$ be the aforementioned M-point approximation and let $\gamma^* \in \Gamma(\mu^M, \mathscr{M}(f))$ be the optimal transference plan. We have

$$\rho_F(\mu^M, f) = \int_{[0,1]^2} |x-y| d\gamma^* = \sum_{i=1}^M \int_{[0,1]} |x-x_i| d\gamma^*(\{x_i\}, \cdot) =$$

$$= \sum_{i=1}^M \int_{[a_i, a_{i+1}]} |x-x_i| d\gamma^*(\{x_i\}, \cdot) = \sum_{i=1}^M \int_{[a_i, a_{i+1}]} |x-x_i| d\gamma^*([0,1], \cdot) =$$

$$= \sum_{i=1}^M \int_{[a_i, a_{i+1}]} f(x)|x-x_i| dx = \sum_{i=1}^M \rho_F(\mu^M|_{[a_i,a_{i+1}]}, f|_{[a_i,a_{i+1}]}).$$

Since each term can be optimized independently, $m_i \delta_{x_i}$ is the optimal 1-point approximation of $\mathscr{M}(f)|_{[a_i,a_{i+1}]}$ and therefore, according to Lemma 46, $x_i = x^*_{[a_i,a_{i+1}]}$ is the central point of measure $\mathscr{M}(f)|_{[a_i,a_{i+1}]}$. □

Definition 60. Let $M \in \mathbb{N}$ be a fixed natural number, and let $\{[a_i, a_{i+1}]\}_{i=1}^M$ be a partition of the interval $[0,1]$. Given a positive function $f \in C[0,1]$, we define an operator $\Phi : [0,1]^{M-1} \to [0,1]^M$

$$X(a_2, ..., a_M) = (x^*_{[0,a_2]}, x^*_{[a_2,a_3]}, ..., x^*_{[a_{M-1},a_M]}, x^*_{[a_M,1]})$$

and an operator $\Psi : [0,1]^M \to [0,1]^{M-1}$

$$A(x_1, x_2, ..., x_M) = \left(\frac{x_1+x_2}{2}, \frac{x_2+x_3}{2}, ..., \frac{x_{M-1}+x_M}{2} \right).$$

It is clear that the optimal approximation of $\mathscr{M}(f)$ is uniquely defined by a partition of $[0,1]$ and that each partition uniquely defines a candidate for an optimal approximation by Corollary 59. The problem of finding optimal approximation is therefore reduced to finding an optimal partition $\mathbf{a} \in \mathbb{R}^{M-1}$. The main tool used in [68] is based on the observation that $A(X(\mathbf{a}))$ provides a better approximation than $\mathbf{a}$, and consequently necessary and sufficient conditions for the sequence $((A \circ X)^n(\mathbf{a}))_{n=1}^\infty$ to converge to the optimal partition are found. In the next part of this section we introduce a method for improving the convergence rate of the optimization process.

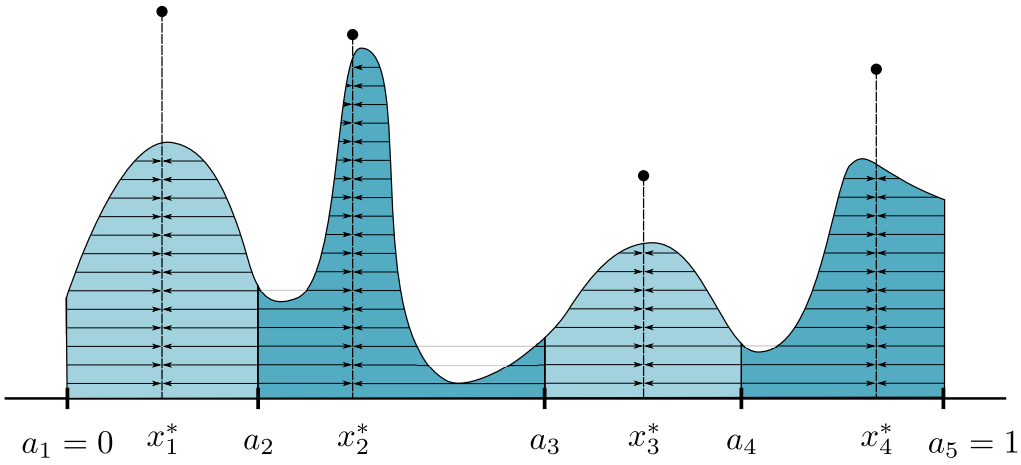
Definition 61. Let $f \in C[0,1]$ and let operator $\omega : [0,1]^{M+1} \to \mathfrak{M}[0,1]$ transforms any partition, $\{a_i\}_{i=1}^{M+1}$, of the interval $[0,1]$ into the optimal M-point approximation in the class of measures whose transport partition coincides with $\{a_i\}_{i=1}^{M+1}$. We shall often write $\mu^M_{\{a_i\}}$ instead of $\omega(\{a_i\})$ for simplicity.

Definition 62. Let $\{a_i\}_{i=1}^{M+1}$ be a fixed partition of the interval $[0,1]$. We define a mapping $\mu : \{2,3,...,M\} \times [0,1] \to \mathfrak{M}[0,1]$ by

$$\mu_{j,a} = \omega(a_1, a_2, ..., a_{j-1}, a, a_{j+1}, ...a_M, a_{M+1}).$$

In other words $\mu_{j,a}$ is the value of ω at the point $\{a_i\}_{i=1}^{M+1}$ with a_j substituted with $a \in [a_{j-1}, a_{j+1}]$.

Figure 1.6: An example of continuous function, a four-point discrete approximation depicted as black dots, the corresponding transport partition, $\{a_i\}_{i=1}^5$ of interval $[0,1]$ and optimal transference plan depicted as horizontal arrows.



Notice that

$$\mu_{i,a} = \sum_{j=1}^{M} m_j^{i,a} \delta_{x_j^{i,a}},$$

where

$$m_j^{i,a} = \begin{cases} \int_{a_j}^{a_{j+1}} f(x)dx & \text{for } j \notin \{i-1, i\} \\ \int_{a_{i-1}}^{a} f(x)dx & \text{for } j = i-1 \\ \int_{a}^{a_{i+1}} f(x)dx & \text{for } j = i \end{cases} \quad \text{and} \quad x_j^{i,a} = \begin{cases} x^*_{[a_j, a_{j+1}]} & \text{for } j \notin \{i-1, i\} \\ x^*_{[a_{i-1}, a]} & \text{for } j = i-1 \\ x^*_{[a, a_{i+1}]} & \text{for } j = i \end{cases}.$$

Theorem 63. *Let $\{a_i\}_{i=1}^{M+1}$ be any increasing sequence on $[0,1]$ with $a_1 = 0$ and $a_{M+1} = 1$. Let $f \in C[0,1]$ be a positive function then for every $i = 2, 3, ..., M$ it holds*

$$\frac{d}{da}\rho_F(\mu_{i,a}, f)\Big|_{a=a_i} = f(a_i)\left[(a_i - x_{i-1}) - (x_i - a_i)\right].$$

Proof. Define $\rho : (a_{i-1}, a_{i+1}) \to \mathbb{R}$ as

$$\rho(a) = \rho_F(\mu_{i,a}, f).$$

To prove the theorem we show that ρ is differentiable by computing the limit

$$\rho'(a) = \lim_{h \to 0} \frac{\rho(a+h) - \rho(a)}{h}.$$

Let $\mu_{i,a} = \sum_{j=1}^{M} m_j^{i,a} \delta_{x_j^{i,a}}$. From Theorem 11 we obtain

$$\rho(a) = \int_0^1 \left| \int_0^t f(\tau)d\tau - \mu_{i,a}[0,t] \right| dt =$$

$$= \int_{a_{i-1}}^{x_{i-1}^{i,a}} \int_{a_{i-1}}^{t} f(\tau)d\tau dt + \int_{x_{i-1}^{i,a}}^{a} \int_{t}^{a} f(\tau)d\tau dt + \int_{a}^{x_i^{i,a}} \int_{a}^{t} f(\tau)d\tau dt + \int_{x_i^{i,a}}^{a_{i+1}} \int_{t}^{a_{i+1}} f(\tau)d\tau dt +$$
$$+ \int_{[0,a_{i-1}]\cup[a_{i+1},1]} \left| \int_0^t f(\tau)d\tau - \mu_{i,a}[0,t] \right| dt.$$

From Corollary 59 and the definition of central point of measure we obtain

$$\int_{a_{i-1}}^{x_{i-1}^{i,a}} f(t)dt = \int_{x_{i-1}^{i,a}}^{a} f(t)dt \quad \text{and} \quad \int_{a_{i-1}}^{x_{i-1}^{i,a+h}} f(t)dt = \int_{x_{i-1}^{i,a+h}}^{a+h} f(t)dt,$$

so by subtracting the left-hand side equation from the right-hand side one we get

$$\int_{x_{i-1}^{i,a}}^{x_{i-1}^{i,a+h}} f(t)dt = \int_{x_{i-1}^{i,a}}^{x_{i-1}^{i,a+h}} f(t)dt + \int_{a}^{a+h} f(t)dt.$$

Consequently,

$$\int_{x_{i-1}^{i,a}}^{x_{i-1}^{i,a+h}} f(t)dt = \int_{x_i^{i,a}}^{x_i^{i,a+h}} f(t)dt = \frac{1}{2}\int_{a}^{a+h} f(\tau)d\tau.$$

Hence,

$$|x_i^{i,a+h} - x_i^{i,a}| \leq h\frac{\sup_{t\in[0,1]}\{f(t)\}}{2\inf_{t\in[0,1]}\{f(t)\}}.$$

We compute

$$\rho(a+h)-\rho(a) = \int_{a_{i-1}}^{x_{i-1}^{i,a+h}} \int_{a_{i-1}}^{t} f(\tau)d\tau dt + \int_{x_{i-1}^{i,a+h}}^{a+h} \int_{t}^{a+h} f(\tau)d\tau dt + \int_{a+h}^{x_i^{i,a+h}} \int_{a+h}^{t} f(\tau)d\tau dt +$$
$$+ \int_{x_i^{i,a+h}}^{a_{i+1}} \int_{t}^{a_{i+1}} f(\tau)d\tau dt - \int_{a_{i-1}}^{x_{i-1}^{i,a}} \int_{a_{i-1}}^{t} f(\tau)d\tau dt - \int_{x_{i-1}^{i,a}}^{a} \int_{t}^{a} f(\tau)d\tau dt +$$
$$- \int_{a}^{x_i^{i,a}} \int_{a}^{t} f(\tau)d\tau dt - \int_{x_i^{i,a}}^{a_{i+1}} \int_{t}^{a_{i+1}} f(\tau)d\tau dt =$$
$$= I_1 + I_2 + I_3 + I_4 - I_5 - I_6 - I_7 - I_8 =$$

$$= \int_{x_{i-1}^{i,a}}^{x_{i-1}^{i,a+h}} \int_{a_{i-1}}^{t} f(\tau)d\tau dt - \int_{x_{i-1}^{i,a}}^{x_{i-1}^{i,a+h}} \int_{t}^{a} f(\tau)d\tau dt + \int_{x_i^{i,a}}^{x_i^{i,a+h}} \int_{a+h}^{t} f(\tau)d\tau dt - \int_{x_i^{i,a}}^{x_i^{i,a+h}} \int_{t}^{a_{i+1}} f(\tau)d\tau dt +$$
$$+ \int_{a}^{a+h} \int_{t}^{a+h} f(\tau)d\tau dt - \int_{a}^{a+h} \int_{a}^{t} f(\tau)d\tau dt + \left[\left(a - x_{i-1}^{i,a+h}\right) - \left(x_i^{i,a} - a\right)\right] \int_{a}^{a+h} f(\tau)d\tau =$$

$$= (I_1 - I_5) + (I_2 - I_6)_1 + (I_3 - I_5)_2 + (I_4 - I_8) + (I_2 - I_6)_2 + (I_3 - I_5)_2 + ((I_2 - I_6) + (I_3 - I_5))_3 .$$

Notice that all the terms apart from $((I_2 - I_6) + (I_3 - I_5))_3$ are of order $\mathcal{O}(h^2)$ since $|x_i^{i,a+h} - x_i^{i,a}| = \mathcal{O}(h)$. On the other hand,

$$\lim_{h\to 0} \frac{1}{h} \int_a^{a+h} f(\tau) d\tau = f(a),$$

hence,

$$\rho'(a) = f(a) \left[\left(a - x_{i-1}^{i,a}\right) - \left(x_i^{i,a} - a\right) \right].$$

□

Corollary 64. *Let $\{a_i\}_{i=1}^{M+1}$ be the transport partition corresponding to an optimal M-point approximation, $\mu^M = \sum_{i=1}^M m_i \delta_{x_i^*}$, of a positive continuous function $f \in C[0,1]$. Then for every $i = 2, 3, ..., M$ it holds that*

$$4 \geq \frac{f(a_i)}{f(x_{i-1}^*)} + \frac{f(a_i)}{f(x_i^*)}. \tag{1.14}$$

Proof. Theorem 63 guarantees that $a \mapsto \rho_F(\mu^{i,a}, f)$ is a differentiable function and

$$\frac{d}{da}\rho_F(\mu^{i,a}, f) = f(a) \left[\left(a - x_{i-1}^{i,a}\right) - \left(x_i^{i,a} - a\right) \right].$$

By proposition 57 we have $\left(a_i - x_{i-1}^{i,a_i}\right) - \left(x_i^{i,a_i} - a_i\right) = 0$ and hence

$$\frac{d^2}{da^2}\rho_F(\mu^{i,a}, f)\bigg|_{a=a_i} = \lim_{h\to 0} \frac{f(a_i + h)}{h} \left[\left(a_i + h - x_{i-1}^{i,a_i+h}\right) - \left(x_i^{i,a_i+h} - a_i - h\right) \right].$$

Since

$$\frac{d}{da} x_i^{i,a} = \lim_{h\to 0} \frac{x_i^{i,a+h} - x_i^{i,a}}{h} = \frac{1}{2} \frac{f(x)}{f(x_i^{i,a})}$$

we obtain

$$\frac{d^2}{da^2}\rho_F(\mu^{i,a}, f)\bigg|_{a=a_i} = f(a_i) \left[2 - \frac{1}{2} \frac{f(a_i)}{f(x_i^{i,a_i})} - \frac{1}{2} \frac{f(a_i)}{f(x_{i-1}^{i,a_i})} \right].$$

Since $\{a_i\}_{i=1}^{M+1}$ corresponds to the optimal approximation of f we have

$$\frac{d}{da}\rho_F(\mu_{i,a}, f)\bigg|_{a=a_i} = 0$$

and

$$\frac{d^2}{da^2}\rho_F(\mu^{i,a}, f)\bigg|_{a=a_i} \geq 0,$$

which proves the corollary. □

The following Corollary follows directly from Theorem 63.

Corollary 65. *Let $f \in C^1[0,1]$ be a positive function and M be a fixed natural number, then*

$$\begin{aligned}
\frac{d^2}{da_i^2}\rho_F(\mu^M_{\{a_i\}_{i=1}^{M+1}}, f) &= f'(a_i)\left[\left(a_i - x^*_{[i-1,i]}\right) - \left(x^*_{[i,i+1]} - a_i\right)\right] + \\
&\quad + f(a_i)\left[2 - \frac{f(a_i)}{2f(x^*_{[i-1,i]})} - \frac{f(a_i)}{2f(x^*_{[i,i+1]})}\right], \\
\frac{d^2}{da_j da_{j-1}}\rho_F(\mu^M_{\{a_i\}_{i=1}^{M+1}}, f) &= -\frac{f(a_{j-1})f(a_j)}{2f(x^*_{[j-1,j]})}, \\
\frac{d^2}{da_j da_k}\rho_F(\mu^M_{\{a_i\}_{i=1}^{M+1}}, f) &= 0 \qquad \text{for } k \notin \{j-1, j, j+1\}.
\end{aligned}$$

Remark 66. Theorem 63 and Corollary 65 allow application of Newton's optimization algorithm for finding the optimal partition. Starting from any $\mathbf{a}^0 \in [0,1]^{M-1}$ sufficiently close to a local minimum, the Newton's method provides a sequence $\{\mathbf{a}^n\}_{n=1}^{\infty}$ converging to the minimum with quadratic rate.

The following proposition shows that the fixed point of $A \circ X$ is not necessarily unique. Consequently, in the general case, neither the Newton's method nor the iterative method has to converge to a global minimum.

Proposition 67. *There exists a positive continuous function f for which function*

$$\{a_2, a_3, ..., a_M\} \mapsto \rho_F(\mu^M_{\{0,a_2,..,a_M,1\}}, f)$$

has more than one local minimum.

Proof. Consider a positive function $f \in C^1[0,1]$, denote $\int_{(i-1)/N}^{i/N} f = f_i$ and suppose that for $N = 7$ we have $(f_i)_{i=1}^7 = (2,1,1,2,3,1,4)$. It is easy to check that in the class of 2-point approximations both partitions $\{a_i\}_{i=1}^3 = \{0, \frac{3}{7}, 1\}$ and $\{a'_i\}_{i=1}^3 = \{0, \frac{4}{7}, 1\}$ satisfy

$$a_2 - x^*_{[a_1,a_2]} = x^*_{[a_2,a_3]} - a_2,$$

which by Theorem 63 implies that both partitions are the extremum points of error function. Since f is an arbitrary function it can be chosen so that the Hessian, defined as 65, is positive defined. □

Corollary 68. *Let f be a positive continuous function. A local minimum, $\mathbf{a} \in \mathbb{R}^{M-1}$ of function $\{a_2, a_3, ..., a_M\} \mapsto \rho_F(\mu^M_{\{0,a_2,..,a_M,1\}}, f)$ is a fixed-point of operator $A \circ X$.*

Proof. Let $\mathbf{x} = X(\mathbf{a})$. Since $\mathbf{a}$ is a local minimum, then all partial derivatives are equal to 0. Theorem 63 guarantees that $(a_i - x_{i-1}) - (x_i - a_i)$, hence $\mathbf{a} = A(x)$. □

Together with Theorem 53 the following Proposition shows that very smooth functions with low oscillation and measures with uniformly distributed atoms are those that are the hardest to approximate with discrete measures.

Proposition 69. *The optimal approximation of a constant function $f(x) = C$ on $[a, b]$ has the error equal to*

$$\frac{C(b-a)^2}{4N}.$$

Proof. Let $\{a_i\}_{i=1}^{M+1}$ be the partition corresponding to the optimal approximation of f. Since $x^*_{[a_i,a_{i+1}]} = \frac{a_i+a_{i1}}{2}$ from Theorem 11 we conclude that the contribution to the error from each interval $[a_i, a_{i+1}]$ equals

$$2\int_{a_i}^{(a_i+a_{i+1})/2} \int_{a_i}^{t} C d\tau dt = 2C \cdot \int_0^{(a_i+a_{i+1})/2-a_i} t dt = \frac{C}{4}\left(a_{i+1} - a_i\right)^2,$$

so the total error of approximation is given by

$$\frac{C}{4}\sum_{i=1}^{N}(a_{i+1} - a_i)^2.$$

This value is minimized for equidistant partition points a_i, for which the error of the approximation equals

$$\frac{C\left(b-a\right)^2}{4N}.$$

□

Discrete approximations in general cannot guarantee an error of better order than N^{-1}. For some applications it is desirable to approximate functions with a different class measure to obtain lower error. The following theorems demonstrate advantages of approximation by N-step functions (linear combinations of N indicator functions).

Theorem 70. *For every Lipschitz continuous function f there exists an N-step approximation f^N such that*

$$\rho(f, f^N) \leq \frac{Lip(f)}{6} \cdot N^{-2}.$$

Proof. Let f^N be given by

$$f^N = \sum_{i=0}^{N-1} \left(\int_{\frac{i}{N}}^{\frac{i+1}{N}} f(x)dx \right) \mathbb{1}_{[\frac{i}{N}, \frac{i+1}{N}]},$$

then by Lemma 20 and Lemma 39

$$\rho_F(f, g^N) \leq \sum_{i=0}^{N-1} W(f|_{[\frac{i}{N}, \frac{i+1}{N}]}, f^N|_{[\frac{i}{N}, \frac{i+1}{N}]}).$$

By Theorem 11

$$\rho_F(f, f^N) \leq \sum_{i=0}^{N-1} \int_{\frac{i}{N}}^{\frac{i+1}{N}} \left| \int_{\frac{i}{N}}^{x} f(t)dt - \int_{\frac{i}{N}}^{x} f^N(t)dt \right| dx \leq \sum_{i=0}^{N-1} \int_{\frac{i}{N}}^{\frac{i+1}{N}} \int_{\frac{i}{N}}^{x} \left| f(t) - f^N(t) \right| dtdx.$$

By the mean value theorem for each $i \in \{0, 1, .., N-1\}$ there exists t_i, such that $f(t_i) = f^N(t_i)$, hence

$$\rho_F(f, f^N) \leq \sum_{i=0}^{N-1} \int_{\frac{i}{N}}^{\frac{i+1}{N}} \int_{\frac{i}{N}}^{x} Lip(f) \cdot \left(t - \frac{i}{N}\right) dtdx \leq \sum_{i=0}^{N-1} Lip(f) \cdot \frac{N^{-3}}{6} = \frac{Lip(f)}{6} N^{-2}.$$

□

Chapter 2

McKendrick-von Foerster equation

In the original McKendrick-von Foerster model the evolution of an age-structured population is described by a hyperbolic partial differential equation in which time and age are the independent variables, see [56]. McKendrick-von Foerster equation with nonlinear growth, reproduction and mortality rates was studied in the framework of L^p spaces in [2], where convergence of a finite-difference scheme was proved. In this approach, however, it was necessary to make strong assumptions on parameters (e.g. growth rate needs to be twice continuously differentiable with respect to structural variable) and the finite-difference scheme has some undesired properties, such as a wrong propagation speed. In [48] a numerical scheme, based on discontinuous Galerkin method, was proposed to address problems in which parameters are only piecewise regular.

In this chapter we consider measure-valued solutions to a McKendrick-von Foerster system [56], which describes the dynamics of a size-structured populations with nonlinear growth, reproduction and mortality rates. The framework of measure-valued solutions is natural and beneficial for the following main reasons:

1. Singularities in a size-structured population dynamics system are inherent. Under low predation, for instance, individuals reach their maximum size with positive probability, which in terms of population size-distribution can be expressed as a Dirac mass at the upper end point of the size range.

2. Measurements in experimental setups are always discrete, hence any comparison between mathematical models and empirical evidence requires tools for comparing general distributions. Metrics from function spaces, such as L^p norms, may induce misleading results in the case of high population concentration and low accuracy of measurements.

3. The notion of a cohort of individuals and its development in time can be formalized.

4. The ability of solving the system for discrete measures is a basis for efficient and highly parallelizable algorithms such as EBT (see Section 2.2).

2.1. Preliminaries

In order to generalize McKendrick-von Foerster model and define measure-valued solutions it is necessary to find an appropriate metric space. In the case of function-valued solution the obvious choice, namely $L^p(X)$, is a complete Banach space, which allows a range of methods to be used for the analysis. In contrast, the natural choice for measure-valued solutions, namely $(\mathfrak{M}(X), \rho_F)$ is not complete and its Banach completion consists of objects that are difficult to interpret in terms of population distributions. In the following considerations we focus on the case of $X \subseteq \mathbb{R}^d$ and present facts that support the choice of $(\mathfrak{M}^+(X), \rho_F)$ as the space of states for the model.

Proposition 71. *Norms* $\|\cdot\|$ and $\|\cdot\|_F$ are not equivalent on $\mathfrak{M}(X)$.

Proof. Consider sequence $\mu_n = \delta_{n^{-1}}$. We have that

$$2 = \|\mu_n - \delta_0\| \geq \|\mu_n - \delta_0\|_F = n^{-1} \to 0.$$

Consequently, $\mu_n \to \delta_0$ in $\|\cdot\|_F$, but not in $\|\cdot\|$. □

Proposition 72. *The space* $(\mathfrak{M}(X), \rho_F)$ *is not complete.*

Proof. Since $(\mathfrak{M}(X), \|\cdot\|)$ is complete and $\|\cdot\|_F$ is not equivalent to $\|\cdot\|$, the space $(\mathfrak{M}(X), \rho_F)$ cannot be complete. □

Example 73. An example of an object from Banach completion of $(\mathfrak{M}(X), \rho_F)$ that is not in $\mathfrak{M}(X)$ can be constructed as follows:

Let $\mu_n = \sum_{k=1}^{n} \delta_{2^{-k}} - n\delta_0$. For $n \leq m$ we have that

$$\rho_F(\mu_n, \mu_m) = W\left(\sum_{k=n+1}^{m} \delta_{2^{-k}}, (m-n)\,\delta_0\right) = \sum_{k=n+1}^{m} 2^{-k} \leq 2^{-n}.$$

Therefore, μ_n is a Cauchy sequence. It's easy to check that no measure $\mu \in \mathfrak{M}(X)$ is a limit of μ_n.

For the proof of the following proposition we refer to [80].

Proposition 74. *The space* $(\mathfrak{M}^+(X), \rho_F)$ *is complete and separable.*

The measure-valued model of McKendrick-von Foerster is considered in the space $(\mathfrak{M}^+(X), \rho_F)$. Hence, the prediction of population dynamics in time is considered as a function of time, $[0, T]$, with values in $\mathfrak{M}^+(X)$. Model parameters, which define growth, mortality and birth processes, are given by functions $g, m, \beta : [0, T] \times \mathfrak{M}^+(X) \to C^{0,1}(X)$ respectively. Values $g(t, \mu)(s)$, $m(t, \mu)(s)$ and $\beta(t, \mu)(s)$ are interpreted as individual growth rate, mortality rate and reproduction rate of an individual of size s, belonging to a population with structure μ at a time point t.

We restrict our farther considerations to $X = [s_{min}, s_{max}]$. Without loss of generality we assume $s_{min} = 0$. Presented results can be generalized to the case of $X = [s_{min}, \infty)$. It is, however, beyond the scope of this thesis.

Definition 75. By McKendrick-von Foerster model of size-structured population we understand the system

$$\begin{cases} \partial_t u + \partial_s(g(t,u)u) + m(t,u)u = 0 & \text{for } t \in T \\ g(t,u)(0)\left(D_{\mathcal{L}_{\mathbb{R}}} u(t)\right)(0) = \int_0^{s_{max}} \beta(t,u)(s)u(ds) & \\ u(0) = u_0 \in \mathfrak{M}^+[0,s_{max}] & , \end{cases} \tag{2.1}$$

where $D_{\mathcal{L}_{\mathbb{R}}}$ denotes Radon-Nikodym derivative with respect to Lebesgue measure on $\mathbb{R}$.

We investigate solutions $u : [0,T] \to \mathfrak{M}^+[0,s_{max}]$ under following conditions on parameters:

Condition 76. Assume

1. $g, m, \beta \in C_b^{0,1}\left([0,T] \times \mathfrak{M}^+[0,s_{max}]; C^{0,1}[0,s_{max}]\right)$,
2. for every $s \in [0,s_{max})$ it holds that $g(t,u)(s) > 0$ and $g(t,u)(s_{max}) = 0$.

Notation 77. For a given function $f \in C_b^{0,1}\left([0,T] \times \mathfrak{M}^+[0,s_{max}]; C^{0,1}[0,s_{max}]\right)$ denote

$$\|f\|_P = \sup_{\substack{\mu \in \mathfrak{M}^+[0,s_{max}] \\ t \in [0,T]}} \|f(t,\mu)\|_{C^{0,1}[0,s_{max}]} + \sup_{t\in[0,T]} Lip\left(f(t,\cdot)\right) + \sup_{\mu \in \mathfrak{M}^+[0,s_{max}]} Lip\left(f(\cdot,\mu)\right)$$

Following [35] we introduce the notion of weak solution.

Definition 78. By the weak solution to system 2.1 we mean a weak-* continuous mapping $u : [0,T] \to \mathfrak{M}^+[0,s_{max}]$ such that for every test function $\varphi \in C^1\left([0,T] \times [0,s_{max}]\right)$ it holds that

$$\begin{aligned} \langle u(T), \varphi(T,\cdot)\rangle - \langle u_0, \varphi(0,\cdot)\rangle &= \int_0^T \langle u(t), \varphi(t,0)\beta(t,u(t))\rangle\, dt + \\ &+ \int_0^T \langle u(t), \partial_t\varphi(t,\cdot) + g(t,u(t))\partial_s\varphi(t,\cdot) - m(t,u(t))\varphi(t,\cdot)\rangle\, dt. \end{aligned}$$

Theorem 79. *Suppose functions $g, m, \beta : [0,T] \times \mathfrak{M}^+[0,s_{max}] \to ([0,s_{max}] \to \mathbb{R})$ satisfy Condition 76, then there exists a unique weak solution, $u \in C_b^{0,1}\left([0,T]; \mathfrak{M}^+[0,s_{max}]\right)$ to system (2.1). Moreover,*

1. *For every $0 \le t_1 \le t_2 \le T$ there exist constants C_1, C_2 such that*

$$\rho_F(u(t_1), u(t_2)) \le C_1 e^{C_1(t_2-t_1)} \|u_0\| (t_1 - t_2).$$

2. *Let $u_0, \tilde{u}_0 \in \mathfrak{M}^+[0,s_{max}]$ and $g, \tilde{g}, m, \tilde{m}, \beta, \tilde{\beta}$ satisfy Condition 76. Let $u(t)$ and $\tilde{u}(t)$ solve system (2.1) for parameters (g,m,β) and $(\tilde{g},\tilde{m},\tilde{\beta})$ respectively. There exist constants C_1, C_2, C_3 such that for every $t \in [0,T]$ it holds that*

$$\rho_F(u(t), \tilde{u}(t)) \le e^{C_1 t}\rho_F(u_0, \tilde{u}_0) + C_2 e^{C_3 t} \left\| (g,m,\beta) - (\tilde{g},\tilde{m},\tilde{\beta}) \right\|_{C^{0,1}[0,1]}.$$

For the proof we refer to Theorem 2.13 in [74].

Definition 80. Let (E, ρ) be a metric space. A bounded operator $S : E \times [0, \delta] \times [0, T] \to E$ is called a Lipschitz semiflow if the following conditions are satisfied:

1. $S(0, \tau) = Id$ for $\tau \in [0, T]$,
2. $S(t + s, \tau) = S(t, \tau + s)S(s, \tau)$ for $\tau, s, t \in [0, T]$ such that $\tau + s + t \leq T$,
3. $\rho(S(t, \tau)\mu, S(s, \tau)\nu) \leq L \cdot (\rho(\mu, \nu) + |t - s|)$ for $s, t \in [0, T]$ and some constant L.

The Lipschitz constant of S, $Lip(S)$, is the smallest value of L for which the third condition holds.

The following corollary results from Theorem 79.

Corollary 81. *Suppose functions $g, m, \beta : [0, T] \times \mathfrak{M}^+[0, s_{max}] \to ([0, s_{max}] \to \mathbb{R})$ satisfy Condition 76, and $u(t)$ is the weak solution to system (2.1). There exists a Lipschitz semiflow $S : \mathfrak{M}^+[0, s_{max}] \times [0, T] \times [0, T] \to \mathfrak{M}^+[0, s_{max}]$ such that*

$$S(t_2 - t_1, t_1)u(t_1) = u(t_2)$$

for every $t_1, t_2 \in [0, T]$.

The following proposition, provides a generalization of the characteristic method, for measure-valued solutions. The result is not surprising, but seems to be absent in the literature.

Theorem 82. *Suppose functions $g, m, \beta : [0, T] \times \mathfrak{M}^+[0, s_{max}] \to ([0, s_{max}] \to \mathbb{R})$ satisfy Condition 76, and $u(t)$ is the weak solution to system (2.1). Let $u(t_0)([a_0, b_0]) = n_0$ for some $0 \leq a_0 \leq b_0 \leq s_{max}$, then*

$$u(t)([a(t), b(t)]) = n_0 - \int_{t_0}^{t} \int_{a(t)}^{b(t)} m(\tau, u(\tau))(x) \cdot u(\tau)(dx)d\tau$$

for

$$\begin{aligned} a(t) &= a_0 + \int_{t_0}^{t} g(\tau, u(\tau))(a(\tau))d\tau, \\ b(t) &= b_0 + \int_{t_0}^{t} g(\tau, u(\tau))(b(\tau))d\tau. \end{aligned} \tag{2.2}$$

Proof. For $1 \gg \varepsilon > 0$ choose $\psi_0^\varepsilon \in C^1[0, s_{max}]$ such that

$$\psi_0^\varepsilon(x) = \begin{cases} 1 & \text{if } x \in [a_0, b_0] \\ 0 & \text{if } x \leq a_0 - \varepsilon \text{ or } x \geq b_0 + \varepsilon \end{cases}.$$

Let $\psi^\varepsilon \in C^1([t_0,t_1]\times[0,s_{max}])$ be a solution to

$$\begin{cases} \frac{\partial}{\partial t}\psi^\varepsilon(t,x) + g(t,u(t))(x)\frac{\partial}{\partial x}\psi^\varepsilon(t,x) = 0 \quad \text{on } [t_0,t_1]\times[0,s_{max}] \\ \psi^\varepsilon(t_0,\cdot) = \psi_0^\varepsilon \\ \psi^\varepsilon(\cdot,0) = 0 \end{cases} . \tag{2.3}$$

and for every $x \in [0,s_{max}]$ let $l_x(t)$ satisfy

$$\begin{cases} \frac{d}{dt}l_x(t) = g(t,u(t))(l_x(t)) \\ l_x(t_0) = x \end{cases} .$$

From the usual characteristic method for classical solutions to (2.3) we obtain

$$\psi^\varepsilon(t,l_x(t)) = \psi^\varepsilon(t_0,x).$$

Finally let $\varphi^\varepsilon \in C^1([0,T]\times[0,s_{max}])$, be an extension of ψ^ε satisfying

$$\varphi^\varepsilon(t,x) = \begin{cases} \psi^\varepsilon(t,x) & \text{if } t \in [t_0,t_1] \\ 0 & \text{if } t \le t_0-\varepsilon \text{ or } t \ge t_1+\varepsilon \\ 0 & \text{if } x \le l_{a_0-\varepsilon}(t) \text{ or } x \ge l_{b_0+\varepsilon}(t) \end{cases} .$$

We also require that $|\frac{\partial}{\partial t}\varphi^\varepsilon(t,x)| \le 2\varepsilon^{-1}$ and $|\frac{\partial}{\partial x}\varphi^\varepsilon(t,x)| = 0$ for

$$t \in [t_0-\varepsilon,t_0]\cup[t_1,t_1+\varepsilon].$$

Additionally, we choose $\frac{\partial}{\partial t}\varphi^\varepsilon(t,x)$ to be equal $\frac{1}{\varepsilon}$ and $\frac{-1}{\varepsilon}$ on $[l_{a_0(t_0)},l_{b_0(t_0)}]\times U_0$ and $[l_{a_0(t_1)},l_{b_0(t_1)}]\times U_1$ respectively, where $U_0 \subset [t_0-\varepsilon,t_0]$ and $U_1 \subset [t_1,t_1+\varepsilon]$ are some intervals such that $|U_0| \le \varepsilon(1-\varepsilon)$, $|U_1| \le \varepsilon(1-\varepsilon)$.

For every $t \in [t_0,t_1]$ function $\varphi^\varepsilon(t,\cdot)$ is supported on $[l_{a_0-\varepsilon}(t),l_{b_0+\varepsilon}(t)]$. Since

$$\begin{aligned} \frac{d}{dt}\left(l_{b_0+\varepsilon}(t)-l_{b_0}(t)\right) &= g(t,u(t))\left(l_{b_0+\varepsilon}(t)\right) - g(t,u(t))\left(l_{b_0}(t)\right) \le \\ &\le Lip\left(g(t,u(t))\right)\cdot\left(l_{b_0+\varepsilon}(t)-l_{b_0}(t)\right) \end{aligned}$$

we have that

$$[l_{a_0-\varepsilon}(t),l_{b_0+\varepsilon}(t)] \subseteq [l_{a_0}(t)-C_1\varepsilon, l_{b_0}(t)+C_1\varepsilon]$$

for

$$C_1 = exp\left(T\cdot\sup_{t\in[t_0,t_1]} Lip(g(t,u(t)))\right).$$

Choose $\varepsilon > 0$ such that $0 \le t_0-\varepsilon \le t_1+\varepsilon \le T$. By the definition of weak solution to (2.1) for test function φ^ε we have that

$$0 = \int_0^T \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot) + g(t,u(t))\partial_s\varphi_\varepsilon(t,\cdot)\rangle - \langle u(t), m(t,u(t))\varphi_\varepsilon(t,\cdot)\rangle\, dt,$$

hence

$$\int_{t_0-\varepsilon}^{t_1+\varepsilon} \langle u(t), m(t,u(t))\varphi_\varepsilon(t,\cdot)\rangle\, dt = \int_{t_0}^{t_0+\varepsilon} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot) + g(t,u(t))\partial_s\varphi_\varepsilon(t,\cdot)\rangle\, dt + + \int_{t_1}^{t_1+\varepsilon} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot) + g(t,u(t))\partial_s\varphi_\varepsilon(t,\cdot)\rangle\, dt.$$

By the dominated convergence theorem

$$\lim_{\varepsilon\to 0} \langle u(t), m(t,u(t))\varphi_\varepsilon(t,\cdot)\rangle = \int_0^{s_{max}} m(t,u(t))\mathbb{1}_{[l_{a_0}(t), l_{b_0}(t)]} u(t)(dx),$$

thus

$$\lim_{\varepsilon\to 0}\int_{t_0-\varepsilon}^{t_1+\varepsilon} \langle u(t), m(t,u(t))\varphi_\varepsilon(t,\cdot)\rangle\, dt = \int_{t_0}^{t_1}\int_{l_{a_0(t)}}^{l_{b_0(t)}} m(t,u(t))(x)\cdot u(t)(dx)dt.$$

On the other hand,

$$\int_{t_1}^{t_1+\varepsilon} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot) + g(t,u(t))\partial_s\varphi_\varepsilon(t,\cdot)\rangle\, dt = \int_{t_1}^{t_1+\varepsilon} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot)\rangle\, dt$$

and

$$\lim_{\varepsilon\to 0}\int_{t_1}^{t_1+\varepsilon} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot)\rangle\, dt = \lim_{\varepsilon\to 0}\int_{U_1} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot)\rangle\, dt + \int_{[t_1,t_1+\varepsilon]\setminus U_1} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot)\rangle\, dt =$$

$$= \lim_{\varepsilon\to 0}\frac{-1}{\varepsilon}\int_{t_1}^{t_1+\varepsilon}\int_{l_{a_0(t)}}^{l_{b_0}(t)} u(t)(dx)dt + \int_{[t_1,t_1+\varepsilon]\setminus U_1} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot)\rangle\, dt.$$

Since

$$\left|\int_{[t_1,t_1+\varepsilon]\setminus U_1} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot)\rangle\, dt\right| \le \varepsilon^2 \cdot \sup_{t\in[t_1,t_1+\varepsilon]} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot)\rangle \le \varepsilon^2\cdot 2\varepsilon^{-1}\cdot \sup_{t\in[0,T]} u(t)([0,s_{max}])$$

we have that

$$\lim_{\varepsilon\to 0}\int_{t_1}^{t_1+\varepsilon} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot) + g(t,u(t))\partial_s\varphi_\varepsilon(t,\cdot)\rangle\, dt = -\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\int_{t_1}^{t_1+\varepsilon}\int_{l_{a_0(t)}}^{l_{b_0}(t)} u(t)(dx)dt.$$

From weak-$*$ continuity of u with respect to time variable, we obtain

$$\int_{t_1}^{t_1+\varepsilon} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot) + g(t,u(t))\partial_s\varphi_\varepsilon(t,\cdot)\rangle\, dt \to -\int_{l_{a_0(t_1)}}^{l_{b_0}(t_1)} u(t_1)\,(dx)\,,$$

and by the same arguments

$$\int_{t_0}^{t_0+\varepsilon} \langle u(t), \partial_t\varphi_\varepsilon(t,\cdot) + g(t,u(t))\partial_s\varphi_\varepsilon(t,\cdot)\rangle\, dt \to \int_{l_{a_0(t_0)}}^{l_{b_0}(t_0)} u(t_0)\,(dx)\,.$$

Since $x(t) = l_{x_0}(t)$ we obtain that for every $t_0 \le t_1 < T$ it holds that

$$u(t_1)\,([l_{a_0}(t_1), l_{b_0}(t_1)]) = u(t_0)\,([a_0,b_0]) - \int_{t_0}^{t_1}\int_{l_{a_0(t)}}^{l_{b_0(t)}} m(t,u(t))(x)\cdot u(t)(dx)dt,$$

which completes the proof. □

Proposition 83. *Suppose functions $g, m, \beta : [0,T] \times \mathfrak{M}^+[0, s_{max}] \to ([0, s_{max}] \to \mathbb{R})$ satisfy Condition 76, and $u(t)$ is the weak solution to system (2.1). If $l^1(t)$ is the solution of*

$$\begin{cases} \frac{d}{dt} l^1(t) = g(t, u(t))\,(l^1(t)) \\ l^1(0) = 0 \end{cases},$$

then $u(t_1)$ is absolutely continuous on $[0, l^1(t_1)]$ with respect to the Lebesgue measure.

Proof. It is sufficient to prove that for some constant C and every pair $a, b \in [0, l^1(t_1)]$ it holds that

$$u(t_1)([a,b]) \leq C \cdot |b-a|.$$

Let $l_s(t)$ be the solution of

$$\begin{cases} \frac{d}{dt} l_s(t) = g(t, u(t))\,(l_s(t)) \\ l_s(t_1) = s \end{cases}.$$

Since for every $x \in [0, s_{max})$ the value of $g(t, u(t))(x)$ is strictly positive, then for every $s \in [0, l^1(t_1)]$ there exists an instant of time, $0 \leq t_0(s) \leq t_1$, such that $l_s(t_0(s)) = 0$.

For $1 \gg \varepsilon > 0$ choose $\psi_0^\varepsilon \in C^1[0,T]$ such that

$$\psi_0^\varepsilon(t) = \begin{cases} 1 & \text{if } t \in [t_0(a), t_0(b)] \\ 0 & \text{if } t \leq t_0(a) - \varepsilon \text{ or } t \geq t_0(b) + \varepsilon \end{cases}.$$

Let $\psi^\varepsilon \in C^1([t_0, t_1] \times [0, s_{max}])$ be a solution to

$$\begin{cases} \frac{\partial}{\partial t}\psi^\varepsilon(t,x) + g(t, u(t))(x)\frac{\partial}{\partial x}\psi^\varepsilon(t,x) = 0 \quad \text{on } [0,t_1] \times [0, s_{max}] \\ \psi^\varepsilon(\cdot, 0) = \psi_0^\varepsilon \\ \psi^\varepsilon(0, \cdot) = 0 \end{cases}. \tag{2.4}$$

Let $\varphi^\varepsilon \in C^1([0,T] \times [0, s_{max}])$, be an extension of ψ^ε satisfying

$$\varphi^\varepsilon(t,x) = \begin{cases} \psi^\varepsilon(t,x) & \text{if } t \in [0, t_1] \\ 0 & \text{if } t \geq t_1 + \varepsilon \end{cases}.$$

Similarly as in the proof of 82 we require that $|\frac{\partial}{\partial t}\varphi^\varepsilon(t,x)| \leq 2\varepsilon^{-1}$ and $|\frac{\partial}{\partial x}\varphi^\varepsilon(t,x)| = 0$ for $t \in [t_0 - \varepsilon, t_0] \cup [t_1, t_1 + \varepsilon]$. Additionally, we choose $\frac{\partial}{\partial t}\varphi^\varepsilon(t,x)$ to be equal $\frac{1}{\varepsilon}$ and $\frac{-1}{\varepsilon}$ on $[l_{a_0(t_0)}, l_{b_0(t_0)}] \times U_0$ and $[l_{a_0(t_1)}, l_{b_0(t_1)}] \times U_1$ respectively, where $U_0 \subset [t_0 - \varepsilon, t_0]$ and $U_1 \subset [t_1, t_1 + \varepsilon]$ are some intervals such that $|U_0| \leq \varepsilon(1-\varepsilon)$, $|U_1| \leq \varepsilon(1-\varepsilon)$. By the definition of weak solution to (2.1) for test function φ^ε we have that

$$\begin{aligned} 0 &= \int_0^T \langle u(t), \partial_t \varphi^\varepsilon(t,\cdot) + g(t,u(t))\partial_s \varphi^\varepsilon(t,\cdot) - m(t,u(t))\varphi^\varepsilon(t,\cdot)\rangle\, dt + \\ &+ \int_0^T \langle u(t), \psi_0^\varepsilon(t)\beta(t,u(t))\rangle\, dt. \end{aligned}$$

and consequently

$$0 \leq \int_0^T \langle u(t), \partial_t \varphi^\varepsilon(t,\cdot) + g(t,u(t))\partial_s \varphi^\varepsilon(t,\cdot)\rangle \, dt + \|\beta\|_P \cdot \int_{t_0(a)-\varepsilon}^{t_0(b)+\varepsilon} u(t)([0,s_{max}])dt.$$

Passing with $\varepsilon \to 0$ we obtain

$$u(t_1)([a,b]) \leq \|\beta\|_P \cdot \sup_{t\in[0,T]} u(t)([0,s_{max}]) \cdot |t_0(b) - t_0(a)|,$$

which completes the proof since $t_0(\cdot)$ is a Lipschitz continuous function. □

Definition 84. By a stationary state we mean the value, μ, of a solution $u : [0,T] \to \mathfrak{M}^+[0,s_{max}]$ which is not dependent on time, namely $\mu = u(t)$ for every $t \in [0,T]$.

The following lemma states that even in the general framework of measure-valued solutions all possible stationary states are absolutely continuous under some reasonably weak conditions.

Lemma 85. *Suppose functions $g, m, \beta : [0,T] \times \mathfrak{M}^+[0,s_{max}] \to ([0,s_{max}] \to \mathbb{R})$ satisfy Condition 76, and $\mu \in \mathfrak{M}^+[0,s_{max}]$ is a stationary state of system (2.1). If*

$$m(t,\mu)(s_{max}) > 0$$

for some $t \in [0,T]$, then μ is absolutely continuous with respect to Lebesgue measure.

Proof. Let μ be a stationary state of equation (2.1) and let $l^1(t) : \mathbb{R}^{\geq 0} \to [0,s_{max}]$ be defined as in Theorem 83. Since $g(t,\mu)(x) > 0$ for every $x < s_{max}$ we obtain that

$$\lim_{t\to\infty} l^1(t) = s_{max}.$$

By Theorem 83 solution $u(t)$ to (2.1) is absolutely continuous on $[0, l^1(t)]$. Consequently, the stationary state, $\mu = u(t)$, is absolutely continuous on interval $[0, s_{max})$. It implies that $\mu = \mu_{ac} + m_{s_{max}}\delta_{s_{max}}$, where μ_{ac} is absolutely continuous with respect to Lebesgue measure. By Theorem 82 we obtain that

$$m_{s_{max}} = m_{s_{max}} \left(1 - \int_{t_1}^{t_2} m(\tau,\mu)(s_{max})d\tau\right).$$

Therefore either $m_{s_{max}} = 0$ or $m(t,\mu)\,(s_{max}) = 0$ for all t. □

The following lemma provides a characterization of demographic trends in stationary state.

Lemma 86. *Suppose functions $g, m, \beta : [0,T] \times \mathfrak{M}^+[0,s_{max}] \to ([0,s_{max}] \to \mathbb{R})$ satisfy Condition 76, and $\mu \in \mathfrak{M}^+[0,s_{max}]$ is a stationary state of system (2.1) then for every $t \in [0,T]$ it holds that $\langle \mu, m(t,\mu)\rangle = \langle \mu, \beta(t,\mu)\rangle$.*

Proof. Let u be a weak solution to (2.1). For a test function, being a standard regularization of

$$\varphi(t,s) = \begin{cases} 1 & t \in [t_0, t_1] \\ 0 & \text{otherwise} \end{cases},$$

the definition of weak solution implies

$$u(t_1)([0, s_{max}]) - u(t_0)([0, s_{max}]) = \int_{t_0}^{t_1} \langle u(t), \beta(t, u(t))\rangle \, dt - \int_{t_0}^{t_1} \langle u(t), m(t, u(t))\rangle \, dt.$$

Regularization and passing to the limit is explained in detail in the proof of Theorem 82. Since $u(t_1) = u(t_0) = \mu$ we have that for any $t_0, t_1 \in \mathbb{R}^{\geq 0}$

$$\int_{t_0}^{t_1} \langle \mu, \beta(t, u(t))\rangle \, dt = \int_{t_0}^{t_1} \langle \mu, m(t, u(t))\rangle \, dt,$$

hence, $\langle \mu, \beta(t, u(t))\rangle = \langle \mu, m(t, u(t))\rangle$ for every t. □

2.2. Particle methods

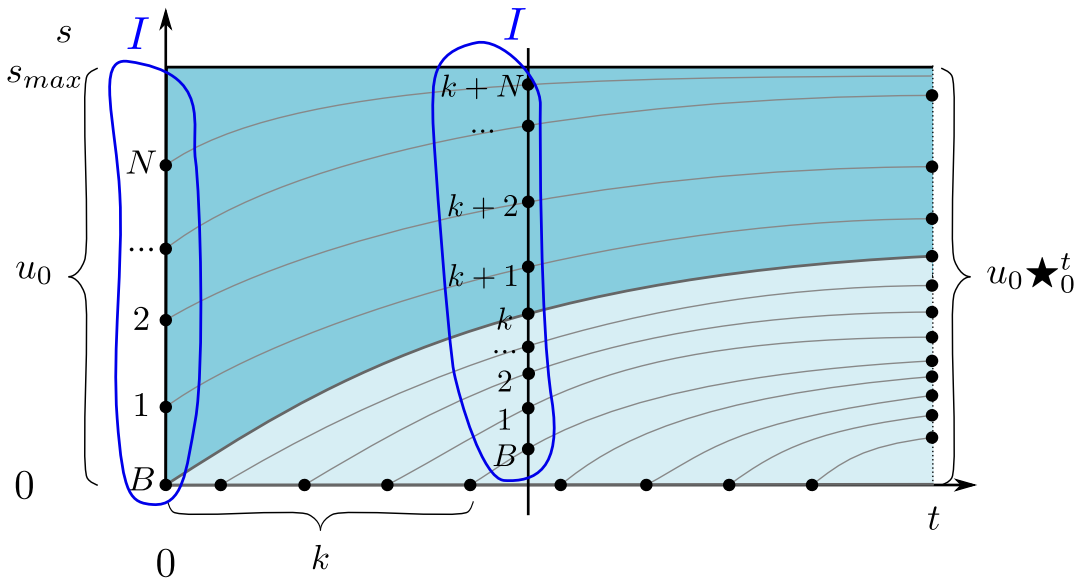
Particle methods is an umbrella term for a wide class of numerical schemes for first order hyperbolic equations. The concept is to approximate the initial conditions by a large number of particles and track each of the particles separately. In this section we focus on Escalator Boxcar Train (EBT) algorithm for solving McKendrick-von Foerster equation with non-local terms reflecting the impact of the whole population on individual birth, growth and death processes. EBT was first introduced in [15] where it was used as a heuristic approach based on the intuition that a continuously distributed population can be studied as a collection of cohorts. Rigorous proof of convergence of this scheme [9] and the analysis of the order of convergence (see [34]) was possible after developing certain tools for the space of measures and Lipschitz semiflows.

In Section 2.2.1 a summary of results from [34] is presented. Author's contribution to this joint paper was limited to simplifying the proofs, implementing the schemes and running numerical tests. Notice that numerical comparison of results requires an implementation of the algorithm described in Section 1.3.4. Three improvements to standard EBT algorithm, which arise from the considerations of Section 1.5, are presented in Section 2.2.3.

2.2.1. EBT algorithm

Particle methods in their principle are based on approximating a solution to partial differential equation by a sum of Dirac masses and tracking each mass in time. The main challenge, as it will become clear after reading this section, is handling the boundary conditions. A number of methods for tracking boundary cohorts has already been developed, and three of them (original EBT, EBT with simplified boundary conditions and Split-Up algorithm) are compared in [34], where no significant differences in the rate of

Figure 2.1: Visualization of the sEBT algorithm



convergence were found. In this section we summarize methods and results used in the analysis of particle algorithms for transport equations with non-local terms.

We restrict our considerations to equation (2.1) with

$$g, m, \beta \in C_b^{0,1}\left([0,T] \times \mathcal{M}^+[0, s_{max}]; C^{0,1}[0, s_{max}]\right)$$

and $u_0 \in \mathfrak{M}[0, s_{max}]$ with possibly infinite s_{max}. We shall also focus on one of the algorithms analyzed in [34], namely on the EBT algorithm with simplified boundary conditions, abbreviated to sEBT. Analysis of other algorithms is very similar and the order of convergence is identical. Numerical results for all three methods are compared in Section 2.2.2.

The main idea of the sEBT method is to approximate the initial conditions $u_0 \in \mathfrak{M}^+[0, s_{max}]$ by a discrete measure $\mu_0 = \sum_{i=1}^{N} m_i(0)\delta_{x_i(0)}$ and "track" position and mass of each Dirac delta (see Figure 2.1). In the case of (2.1) the following ODE system is used for the tracking

$$\begin{cases} \frac{d}{dt}x_i(t) = g(t, \sum_{i\in I} m_i(t)\delta_{x_i(t)})(x_i(t)) \\ \frac{d}{dt}m_i(t) = -m(t, \sum_{i\in I} m_i(t)\delta_{x_i(t)})(x_i(t)) \cdot m_i(t) \end{cases} \tag{2.5}$$

with I being the set of indices. Boundary conditions are dealt with separately. A new boundary cohort is created every $\Delta t > 0$ of time, and the previous boundary cohort becomes an internal cohort tracked by (2.5). Boundary cohorts, on the other hand, are

tracked by the following equation

$$\begin{cases} \frac{d}{dt}x_B(t) = g(t, \sum_{i\in I} m_i(t)\delta_{x_i(t)})(x_B(t)) \\ \frac{d}{dt}m_B(t) = -m(t, \sum_{i\in I} m_i(t)\delta_{x_i(t)})(x_B(t)) \cdot m_i(t) + \sum_{i\in I} \beta(t, \sum_{i\in I} m_i(t)\delta_{x_i(t)})(x_i(t))m_i(t) \\ x_B(k\Delta t) = m_B(k\Delta t) = 0 \end{cases} . \tag{2.6}$$

The set of indices, I, initially consists of the boundary cohort index, B, and N indices of atoms in the initial approximations, see Figure 2.1. Therefore at time $t \in [k\Delta t, (k+1)\Delta t]$ we have

$$I = \{B, 1, 2, ..., k, k+1, k+2, ..., k+N\}.$$

For a given $\mu \in \mathfrak{M}[0, s_{max}]$ and $t_0 \in \mathbb{R}^{\geq 0}$ let v be a weak solution of (2.1) with initial conditions posed by μ at time t_0, namely

$$\begin{cases} \partial_t v + \partial_s(g(t,v)v) + m(t,v)v = 0 \\ g(t,v)(0)\left(D_\lambda v(t)\right)(0) = \int_0^{s_{max}} \beta(t,v)(s)v(ds) \\ v(t_0) = \mu \end{cases} .$$

By Corollary 81 operator

$$\# : \mathfrak{M}^+[0, s_{max}] \times [0,T] \times [0,T] \to \mathfrak{M}^+[0, s_{max}]$$

defined as

$$\mu\#_{t_0}^{t_1} = v(t_1)$$

is a Lipschitz semiflow.

Remark 87. Lipschitz constant of semiflow # depends on T.

Notation 88. We denote the outcome of the sEBT algorithm at t_1 starting from initial conditions μ at t_0 is denoted by $\mu\bigstar_{t_0}^{t_1}$.

Lemma 89. *The outcome of the sEBT algorithm is a Lipschitz continuous measure-valued function, namely for any $\mu \in \mathfrak{M}_d^+[0, s_{max}]$ and $t_0 \in \mathbb{R}^{\geq 0}$ it holds that*

$$\mu\bigstar_{t_0}^{\cdot} \in Lip\left([0,T]; \mathfrak{M}^+[0, s_{max}]\right).$$

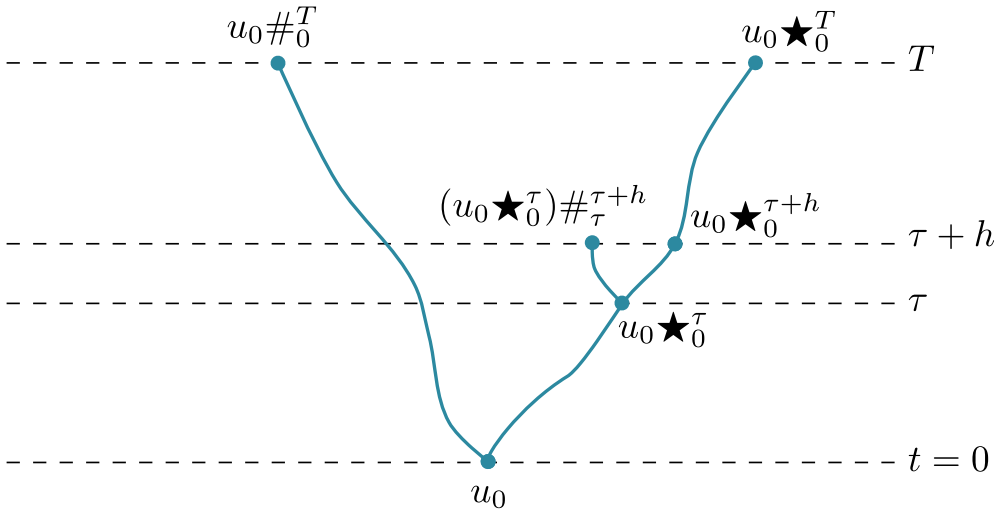
Proof. Lipschitz continuity of functions $x_i(t), m_i(t)$ for $i \in \{B, 1, 2, ..., N\}$ stems from the boundedness of parameters g, m, β. Let $t_1, t_2 \in (t_0, T]$, then

$$\rho_F\left(\mu\bigstar_{t_0}^{t_1}, \mu\bigstar_{t_0}^{t_2}\right) \leq \sum_{i=B}^{N} \rho_F(m_i(t_1)\delta_{x_i(t_1)}, m_i(t_2)\delta_{x_i(t_2)}) + \gamma,$$

where γ is the total amount of newborn individuals added to the boundary cohorts between t_1 and t_2. Consequently, by Lemma 19

$$\rho_F\left(\mu\bigstar_{t_0}^{t_1}, \mu\bigstar_{t_0}^{t_2}\right) \leq \sum_{i=B}^{N} |m_i(t_1) - m_i(t_2)| + \sum_{i=B}^{N} |x_i(t_1) - x_i(t_2)| m_i(t_2) + \gamma \leq$$

Figure 2.2: Corollary 91 provides a method of estimating the error of sEBT scheme on $[0,T]$, namely $\rho_F(u_0\bigstar_0^T, u_0\#_0^T)$, by studying the error on arbitrary small intervals $[\tau, \tau+h]$.



$$\leq |t_1-t_2| \max\left(1, \sum_{i=B}^{N} m_i(t_2)\right)\left(\sup_{i\in\{B,1,\dots,N\}} \|m_i(\cdot)\|_{C^{0,1}[t_0,t_1]} + \sup_{i\in\{B,1,\dots,N\}} \|x_i(\cdot)\|_{C^{0,1}[t_0,t_1]}\right) + \gamma.$$

Finally, by equation (2.6), γ can be estimated by $C(T)\cdot\|\beta\|_P\,|t_1-t_2|$. □

Accuracy of the sEBT algorithm in flat metric, namely $\rho_F(u_0\#_0^T,\, u_0\bigstar_0^T)$, can be estimated from the following theorem (proof can be conducted analogously to the proof of Theorem 2.9 in [10]).

Theorem 90. *Let $S : E\times[0,\delta]\times[0,T]\to E$ be a Lipschitz semiflow. For every Lipschitz continuous map $T:[0,T]\to E$ the following estimate holds*

$$\rho\left(T(t),\, S(t;0)T(0)\right) \leq Lip(S)\cdot\int_0^t \liminf_{h\to 0}\frac{\rho\left(T(\tau+h),\, S(h,\tau)T(\tau)\right)}{h}d\tau.$$

Since $\#$ is a Lipschitz semiflow, Theorem 90 can be applied to the process of population dynamics, $\#$, and the sEBT algorithm, $\bigstar$. The idea hidden behind the following Corollary is depicted on Figure 2.2.

Corollary 91. *Let $u_0\in\mathfrak{M}_d^+[0,s_{max}]$ and $t\in[0,T]$ then*

$$\rho_F(u_0\bigstar_0^t, u_0\#_0^t) \leq tLip(\#)\sup_{\tau\in[0,T]}\liminf_{h\to 0}\frac{\rho_F\left(u_0\bigstar_0^{\tau+h}, (u_0\bigstar_0^\tau)\#_\tau^{\tau+h}\right)}{h}.$$

Theorem 92. *Let $u_0\bigstar_0^\tau = \mu \in \mathfrak{M}_d^+[0, s_{max}]$ be the outcome of sEBT algorithm with time step Δt, then for some constant $C_1(T)$ it holds that*

$$\liminf_{h\to 0} \frac{\rho_F\left(\mu\bigstar_\tau^{\tau+h}, \mu\#_\tau^{\tau+h}\right)}{h} = C_1(T)\Delta t.$$

Proof. Let $\mu = \sum_{i\in I} m_i(\tau)\delta_{x_i(\tau)}$. By Proposition 83 measure $\mu\#_\tau^{\tau+h}$ decomposes to a discrete part $\sum_{i\in I} n_i(\tau+h)\delta_{y_i(\tau+h)}$ and an absolutely continuous measure, $\mathscr{M}\left(f(\tau+h)(\cdot)\right)$. Moreover, for every t function $f(t)(\cdot) \in L^1[0, s_{max}]$ is supported on $[0, l^1(t)]$. By Lemma 20 and Lemma 19 there holds

$$\begin{aligned}
\rho_F\left(\mu\bigstar_\tau^{\tau+h}, \mu\#_\tau^{\tau+h}\right) \;\leq\; & \sum_{i\in I\setminus\{B\}} |m_i(\tau+h) - n_i(\tau+h)| + && (2.7)\\
+ \;& \sum_{i\in I\setminus\{B\}} |x_i(\tau+h) - y_i(\tau+h)| n_i(\tau+h) + && (2.8)\\
+ \;& \rho_F(m_B\delta_{x_B}, n_B\delta_{y_B} + \mathscr{M}\left(f(\tau+h)(\cdot)\right)). && (2.9)
\end{aligned}$$

The first two terms correspond to the error resulting from non-local coefficients b and c. The last term stems from the approximation of a continuous function near the boundary by a 1-point discrete measure. Since the asymptotic behavior of l^1 for $h \to 0$ is given by $l^1(h) = \Theta(h)$ and x_B, y_B may range in $[0, \|g\|_P \Delta t]$, it follows that the contribution of the last term in (2.7) to the total error cannot be estimated from below by a smaller value $\Theta(h\Delta t)$. Indeed, the central point (see Definition 45) of measure $n_B\delta_{y_B} + \mathscr{M}(f(\tau+h)(\cdot))$ tends to y_B with $h \to 0$, and consequently the error of optimal approximation tends to $\Delta t \int_0^{s_{max}} f(\tau+h)(s)ds$. It can be then shown that remaining terms in (2.7) are of order h^2 (see [34] for details). Consequently we obtain

$$\rho_F\left(\mu\bigstar_\tau^{\tau+h}, \mu\#_\tau^{\tau+h}\right) = O(h\Delta t).$$

□

Theorem 93. *Let $u_0 \in \mathfrak{M}^+[0, s_{max}]$ then for any $\mu_0 \in \mathfrak{M}_d^+[0, s_{max}]$ it holds that*

$$\rho_F(\mu_0\bigstar_0^T, u_0\#_0^T) \leq C_1(T)\Delta t + C_2(T)\rho_F(\mu_0, u_0).$$

Proof. Let $\mu_0 \in \mathfrak{M}_d^+[0, s_{max}]$ be an initial approximation of u_0. By Theorem 79 and Corollary 91 we deduce that

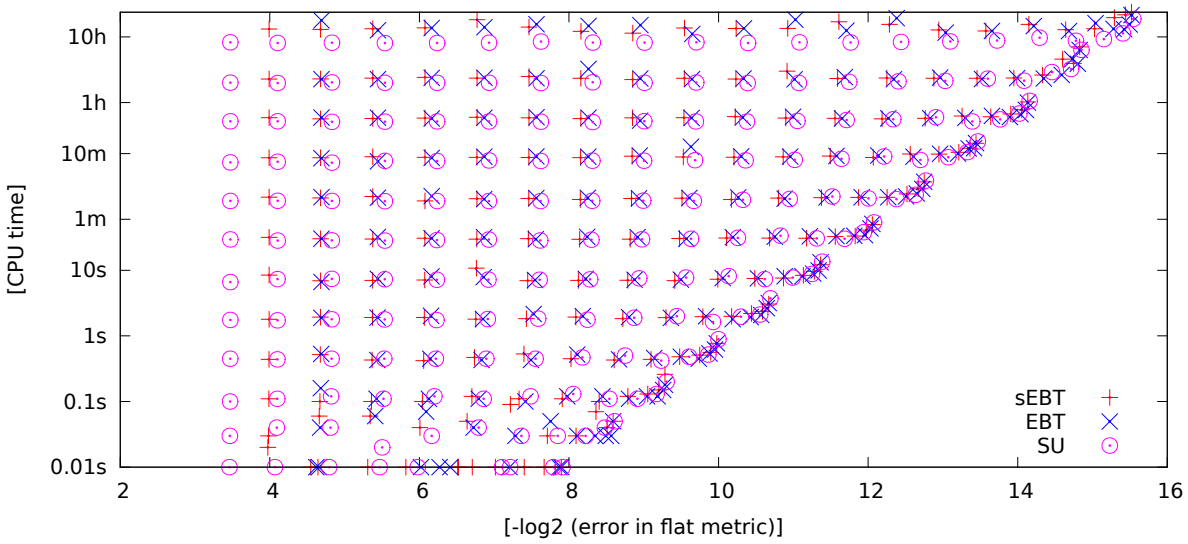
$$\rho_F(\mu_0\bigstar_0^T, u_0\#_0^T) \leq \rho_F(\mu_0\bigstar_0^T, \mu_0\#_0^T) + \rho_F(\mu_0\#_0^T, u_0\#_0^T) \leq C_1(T)\Delta t + C_2(T)\rho_F(\mu_0, u_0).$$

Constant C_2 depends on T since $Lip(\#)$ depends on T. □

2.2.2. Numerical tests

In this section numerical tests of convergence and efficiency of three algorithms described in [34] are presented. The schemes: Escalator Boxcar Train algorithm with simplified

Figure 2.3: CPU time required to achieve given accuracy.



boundary conditions (sEBT), original Escalator Boxcar Train algorithm (EBT) and split-up algorithm (SU) are very similar in essence but differ from each other in the method of handling the boundary cohorts. The results presented in Section 2.2.1 for the sEBT algorithm are easily transferable to the case of SU and EBT.

The numerical simulations show that no major differences in accuracy nor performance are apparent. The tests also confirm the theoretical order of convergence proved in Theorem 93.

The tests were conducted on the following datasets on $X = [0, 1]$:

1. In the first test case we consider a problem with the initial condition taken at a stable stationary state. The aim of the test is to check accuracy of the approximation of the influx modeled by the boundary cohort. We choose the following model parameters:

$$\begin{aligned} g(s) &= 0.2\,(1-s) \\ m(s) &= 0.2 \\ \beta(s) &= 2.4\left(s^2 - s^3\right). \end{aligned}$$

   The exact solution is $u(t) = \mathcal{L}_{[0,1]}$.

2. The second example is taken from the reference [46]. The aim of this test is to study influence of non-local terms on the results for the three algorithms. We take model parameters given by the following functions

$$\begin{aligned} g(s) &= \mathrm{e}^{-s} \\ m(s) &= 1 + \mathrm{e}^{-s} + \frac{\mathrm{e}^{-s}\sin(s)}{2+\cos(s)} \\ \beta(u)(s) &= \frac{3}{2+\cos(s)} \cdot \frac{0.5 + (1 + 0.5\sin(1))\mathrm{e}^{-t}}{0.5 + \langle u, 1\rangle} \end{aligned}$$

The exact solution of the model is $u(t) = \mathrm{e}^{-t}(1 + 0.5\cos(x))\mathcal{L}_{[0,1]}$.

The error of numerical scheme was computed using the algorithm described in Section 1.3.4 as

$$-\log_2\left(\rho_F(\mu(1), u(1))\right),$$

where $\mu(t)$ is the output of the scheme and $u(t)$ is the exact solution.

Figure 2.3 shows the efficiency of EBT algorithms (amount of time required by the central processing unit to obtain desired accuracy). Each point at the plot represents a simulation for a 2^i-point equidistant approximation of initial conditions and $\Delta t = 2^{-k}$ with $i, k \in \{2, 3, ..., 19\}$. The points farthest to the right (high accuracy) correspond to those simulations for which k was close to i. No significant difference in efficiency between algorithms could be found.

Figure 2.4 presents the accuracy of EBT algorithms as a function of the number of initial nodes, I, and boundary cohorts, K. It is clear from the plots that the ratio $1:1$ of initial nodes and boundary cohorts provides the smallest error.

Tables 2.1 and 2.2 provide detailed results and confirm linear order of convergence of the algorithms with respect to Δt. The empirical order of convergence is defined as

$$\log_2 \frac{e_{\frac{I}{2}}}{e_I},$$

where e_I is the error of the numerical scheme for I initial nodes and $\frac{I}{4}$ boundary cohorts.

Table 2.1: Test Case 1. Numerical error and order of convergence measured in flat metric. Number of boundary cohorts equals $I/4$.

	sEBT		EBT		SU	
I	Error	**Order**	Error	**Order**	Error	**Order**
16	1.53e-02	1.03	1.31e-02	1.02	1.49e-02	1.04
32	7.56e-03	1.02	6.56e-03	1.00	7.96e-03	0.90
64	3.76e-03	1.01	3.28e-03	1.00	4.14e-03	0.94
128	1.88e-03	1.00	1.64e-03	1.00	2.11e-03	0.97
256	9.36e-04	1.00	8.20e-04	1.00	1.07e-03	0.99
512	4.68e-04	1.00	4.10e-04	1.00	5.36e-04	0.99
1024	2.34e-04	1.00	2.05e-04	1.00	2.68e-04	1.00
2048	1.17e-04	1.00	1.03e-04	1.00	1.34e-04	1.00
4096	5.84e-05	1.00	5.13e-05	1.00	6.73e-05	1.00
8192	2.92e-05	1.00	2.56e-05	1.00	3.36e-05	1.00
16384	1.46e-05	1.00	1.28e-05	1.00	1.68e-05	1.00
32768	7.30e-06	1.00	6.41e-06	1.00	8.41e-06	1.00
65536	3.65e-06	1.00	3.20e-06	1.00	4.21e-06	1.00
131072	1.83e-06	1.00	1.60e-06	1.00	2.10e-06	1.00
262144	9.13e-07	1.00	8.01e-07	1.00	1.05e-06	1.00
524288	4.56e-07	1.00	4.01e-07	1.00	5.26e-07	1.00
1048576	2.28e-07	1.00	2.00e-07	1.00	2.63e-07	1.00

Figure 2.4: Full map of errors for test case 1 (left) and test case 2 (right) and algorithm sEBT (top), EBT (center), SU (bottom). The plots show the dependence of numerical error in flat metric (Y axis) upon number of initial nodes I (X axis) and the ratio $\frac{K}{I}$ (color). The solid line represents the accuracy of I-point equidistant approximation of the exact solution.

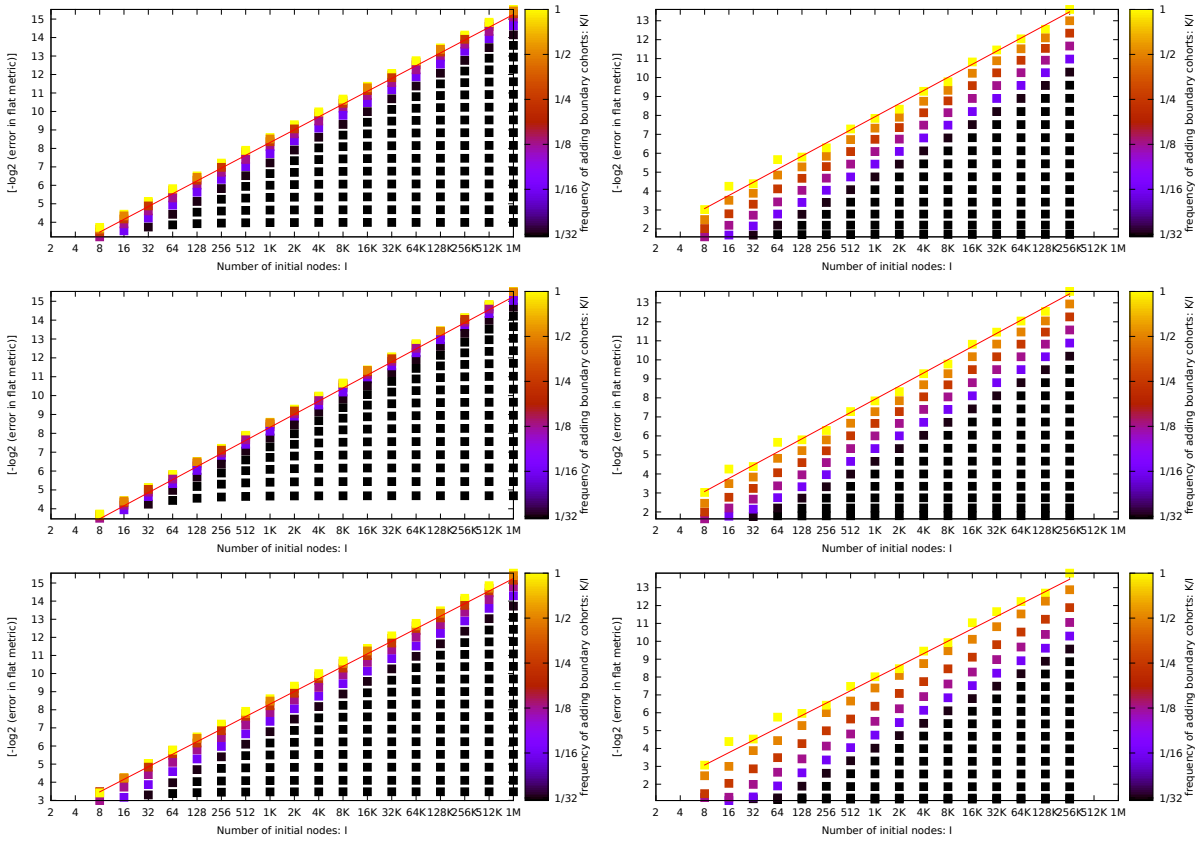


Table 2.2: Test Case 2. Numerical error and order of convergence measured by flat metric. Number of boundary cohorts equals $I/4$.

	sEBT		EBT		SU	
I	Error	**Order**	Error	**Order**	Error	**Order**
16	6.09e-02	1.12	6.25e-02	1.12	1.29e-01	0.82
32	3.67e-02	0.73	3.92e-02	0.67	5.72e-02	1.17
64	1.63e-02	1.17	1.72e-02	1.19	3.06e-02	0.90
128	9.32e-03	0.81	1.01e-02	0.77	1.40e-02	1.13
256	5.02e-03	0.89	5.41e-03	0.90	6.78e-03	1.04
512	2.27e-03	1.15	2.46e-03	1.14	3.52e-03	0.95
1024	1.19e-03	0.93	1.29e-03	0.94	1.72e-03	1.03
2048	6.37e-04	0.90	6.87e-04	0.91	8.42e-04	1.03
4096	2.92e-04	1.12	3.18e-04	1.11	4.33e-04	0.96
8192	1.56e-04	0.91	1.69e-04	0.91	2.12e-04	1.03
16384	6.97e-05	1.16	7.59e-05	1.15	1.11e-04	0.94
32768	3.54e-05	0.98	3.85e-05	0.98	5.48e-05	1.01
65536	1.83e-05	0.95	1.99e-05	0.96	2.70e-05	1.02
131072	9.74e-06	0.91	1.05e-05	0.91	1.32e-05	1.03
262144	4.35e-06	1.16	4.74e-06	1.15	6.91e-06	0.94

2.2.3. Improvements of sEBT algorithm

In this section three improvements to sEBT algorithm, analyzed in Section 2.2.1 and Section 2.2.2, are presented. The first improvement is an application of the theory developed in Section 1.5.3 to reduce the error of initial condition approximation. The second modification makes use of the result of Theorem 51 to reduce complexity of the scheme. Finally, motivated by the result of Theorem 70 we show how the rate of convergence of the sEBT algorithm can be improved if the birth process is approximated by step functions instead of Dirac masses.

2.2.3.1. Initial conditions

Since by Theorem 93 the accuracy of sEBT algorithm is restricted by the time step Δt and the error of the approximation of initial conditions, namely

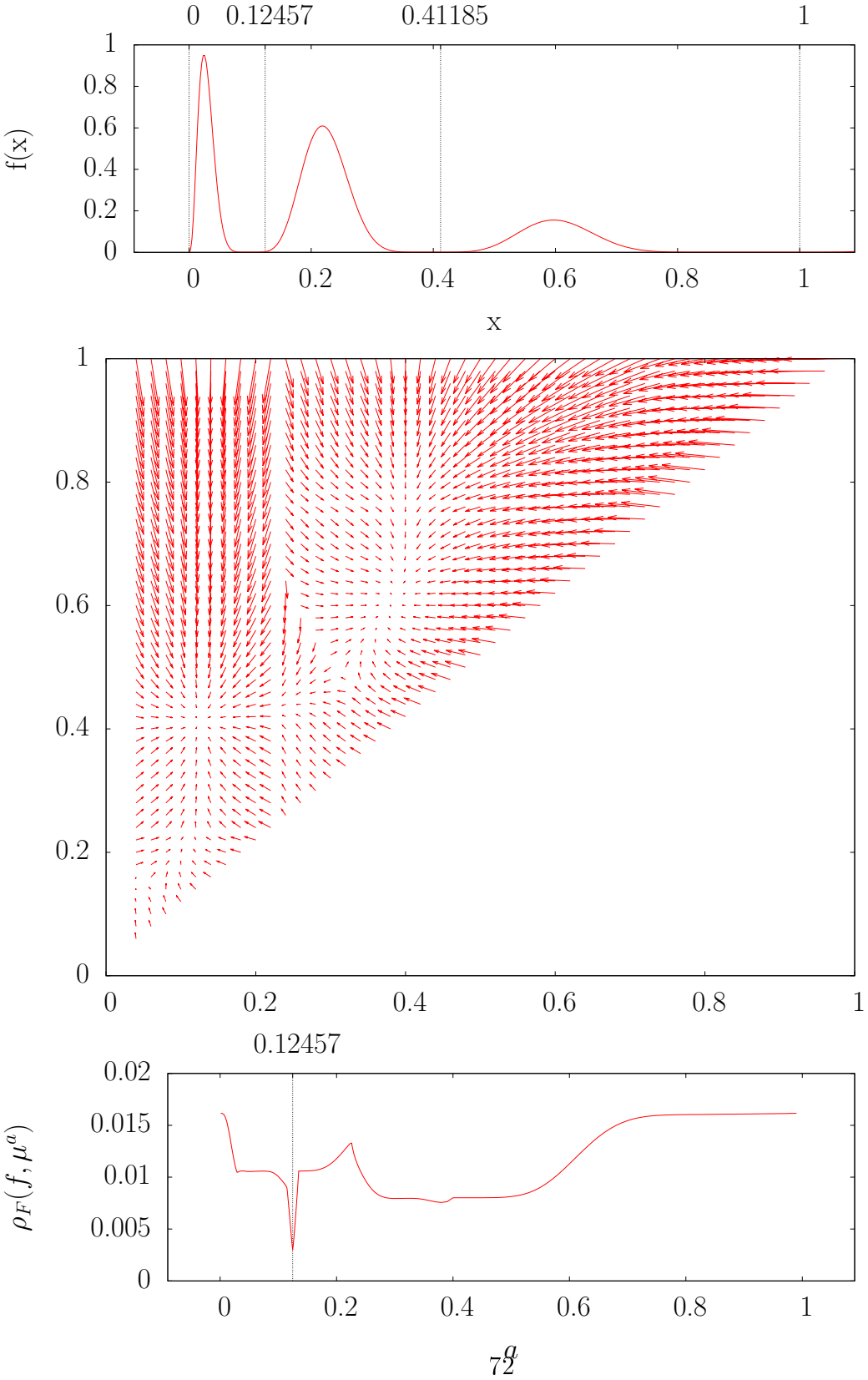
$$\rho_F(\mu_0 \bigstar_0^T, u_0 \#_0^T) \leq C_1(T)\Delta t + C_2(T)\rho_F(\mu_0, u_0),$$

it is natural to apply the results of Section 1.5.3 to reduce the latter factor. From Proposition 69 it is clear, that in the worst case

$$C_2\rho_F(\mu_0, u_0) = \mathcal{O}(\Delta x),$$

where Δx is the maximum distance between two atoms of the initial approximation, μ_0. Yet considerable improvement can be achieved if u_0 is a muliti-hump function.

Figure 2.5: a) function $f(x)$ and its optimal transport partition, b) phase portrait of transport partitions in Newton's method, c) $\rho_F(f, \mu^a)$ as a function of the first non-zero point of transport partition, a_2.



In the farther considerations we will use the following function, $f : [0,1] \to \mathbb{R}^+$, as an illustration for the analyzed tools:

$$f(x) = (1-x)^2 \sin\left(10\sqrt{x}\right)^6 + 10^{-3}.$$

Function f and its optimal transport partition for $N = 3$ are depicted on Figure 2.5a (see Section 1.5.3).

Theorem 63 and Corollary 65 provide sufficient characterization of the error function for Netwon's method to be implemented. Since $a_1 = 0$ and $a_{N+1} = 1$ are fixed, the algorithm finds the minimum argument of the error in $N-1$ dimensional space. For a given point $\mathbf{a}^n \in [0,1]^{N-1}$ Newton method provides a supposedly better point $\mathbf{a}^{n+1} \in [0,1]^{N-1}$, defined as

$$\mathbf{a}^{n+1} = \mathbf{a}^n - \left[H\rho_F(\mu^M_{(0,\mathbf{a}^n,1)}, f)\right]^{-1} \nabla \left[\rho_F(\mu^M_{(0,\mathbf{a}^n,1)}, f)\right],$$

where H denotes the Hessian matrix. Figure 2.5b shows the directions of Netwon steps from different starting points for $N = 3$. Lengths of the arrows, namely $|\mathbf{a}^{n+1} - \mathbf{a}^n|$, were reduced by a factor of 0.2 for clarity.

Another method of finding the optimal approximation is presented on Figure 2.5c. Given $a_2 \in (0,1)$ there exists a unique candidate for the optimal approximation, whose second point of transport partition equals a_2. Indeed, by Corollary 59 the value x_1^* is uniquely defined by a_1 and a_2. Similarly, by Proposition 57 the value a_3 is uniquely defined by x_1^* and a_2. Consequently, given a value a, a transport partition $\{a_i\}_{i=1}^{N+1}$ such that $a_2 = a$ and a corresponding discrete measure, μ^a, can be reconstructed. Figure 2.5c shows the dependence of $\rho_F(f, \mu^a)$ upon a.

The optimal 3-point approximation of function f equals

$$\mu^* = 0.029\delta_{0.027} + 0.055\delta_{0.221} + 0.023\delta_{0.601}$$

and the equidistant 3-point approximation of f equals

$$\mu^3 = 0.084\delta_{\frac{1}{6}} + 0.019\delta_{\frac{1}{2}} + 0.003\delta_{\frac{5}{6}}.$$

Thus,

$$\begin{aligned} \rho_F(\mu^*, f) &= 0.003023, \\ \rho_F(\mu^3, f) &= 0.026841. \end{aligned}$$

2.2.3.2. Reduction of complexity

In sEBT algorithm a new boundary cohort is added every Δt-long period of time. Consequently the number of cohorts grows linearly with time. Assuming that "tracking" a single cohort on an Δt-long interval requires constant computational cost, the algorithm is quadratic with respect to T. Theorem 51 from Section 1.5.2 provides results that allows to reduce a number of cohorts after each time step, and therefore keep it constant.

Proposition 94. *For fixed parameter Δt and fixed approximation of initial condition, μ_0, the computational complexity of sEBT is $\mathcal{O}(T^2)$.*

Proof. Let $\mu_0 \in \mathfrak{M}^+_{d,M}[0, s_{max}]$ and let $N = (\Delta t)^{-1}$. The set of indices, I, defined in Section 2.2.1, at time t_0 has cardinality $M + \lfloor t_0 N \rfloor$. Since tracking a single atom on a time interval $[t_0, t_0 + \Delta t]$ requires a constant computational time, it follows that tracking all particles on the same interval (performing $\bigstar_{t_0}^{t_0+\Delta t}$) requires $\mathcal{O}(M + t_0 \cdot N)$ operations. Consequently, finding approximate solution at time T, namely $\mu_0 \bigstar_0^T$, has computational complexity

$$\mathcal{O}\left(\sum_{k=1}^{NT}(k+M)\right) = \mathcal{O}\left(MNT + N^2T^2\right) = \mathcal{O}(T^2),$$

since

$$\bigstar_0^T = \bigstar_0^{\Delta t} \circ \bigstar_{\Delta t}^{2\Delta t} \circ ... \circ \bigstar_{T-\Delta t}^{T} = \bigcirc_{k=1}^{T\cdot N} \bigstar_{(k-1)\Delta t}^{k\cdot \Delta t}.$$

□

In this section we propose a modification of sEBT which guarantees $\mathcal{O}(T)$ complexity. Throughout the section we assume that $s_{max} \leq 1$.

Definition 95. Let $\blacktriangledown : \mathfrak{M}^+_d(X) \to \mathfrak{M}^+_d(X)$ be a reduction operator which assigns to a measure $\mu \in \mathfrak{M}^+_{d,k}(X)$ its optimal $k-1$-point approximation, $\mu\blacktriangledown$.

We propose a modification of sEBT algorithm in which after each Δt period of time adding new boundary cohort is compensated by optimal reduction by $\blacktriangledown$.

Definition 96. By EBT algorithm with simplified boundary conditions and constant number of cohorts (sEBTc) we mean the following composition:

$$\bigcirc_{k=1}^{T\cdot N}\left(\bigstar_{(k-1)\Delta t}^{k\cdot\Delta t} \circ \blacktriangledown\right) = \bigstar_0^{\Delta t} \circ \blacktriangledown \circ \bigstar_{\Delta t}^{2\Delta t} \circ \blacktriangledown \circ ... \circ \bigstar_{T-\Delta t}^{T},$$

where $N = (\Delta t)^{-1}$.

Proposition 97. *For fixed parameter Δt and fixed approximation of initial conditions, μ_0, the computational complexity of sEBTc is $\mathcal{O}(T)$.*

Proof. Let $\mu_0 \in \mathfrak{M}^+_{d,M}[0, s_{max}]$ and let $N = (\Delta t)^{-1}$. In sEBTc algorithm cardinality of the set of indices, I, is constantly equal $M+1$. Therefore performing $\bigstar_{t_0}^{t_0+\Delta t}$ requires $\mathcal{O}(M)$ operations and by Proposition 50 so does $\blacktriangledown$. Consequently the complexity of the algorithm is given by

$$\mathcal{O}\left(\sum_{k=1}^{NT} M\right) = \mathcal{O}(MNT) = \mathcal{O}(T).$$

□

Theorem 98. *Let $u_0 \in \mathfrak{M}^+[0, s_{max}]$ then for any $\mu_0 \in \mathfrak{M}^+_{d,M}[0, s_{max}]$ and any $\Delta t = N^{-1}$ it holds*

$$\rho_F\left(u_0\#_0^T, \mu_0 \bigcirc_{k=1}^{T\cdot N}\left(\bigstar_{(k-1)\Delta t}^{k\cdot\Delta t} \circ \blacktriangledown\right)\right) \leq C_1(T)\left(\Delta t + NM^{-2}\right) + C_2(T)\cdot\rho_F(\mu_0, u_0)$$

for some constants C_1, C_2 dependent on T.

Proof. Since $\#$ is a Lipschitz semiflow (see Corollary 81) we immediately obtain by triangle inequality

$$\begin{aligned}\rho_F\left(u_0\#_0^T, \mu_0 \bigcirc_{k=1}^{T\cdot N}\left(\bigstar_{(k-1)\Delta t}^{k\cdot\Delta t}\circ\blacktriangledown\right)\right) &\leq Lip(\#)\rho_F(\mu_0,u_0)+\\ &+ \rho_F\left(\mu_0\#_0^T, \mu_0 \bigcirc_{k=1}^{T\cdot N}\left(\bigstar_{(k-1)\Delta t}^{k\cdot\Delta t}\circ\blacktriangledown\right)\right).\end{aligned}$$

It is therefore sufficient to show that the error of sEBTc is of order $\mathcal{O}(\Delta t + NM^{-2})$ if initial condition is a discrete measure. From Theorem 92 and Corollary 91 we have that

$$\rho_F(\mu\bigstar_{t_0}^{t_0+\Delta t}, \mu\#_{t_0}^{t_0+\Delta t}) \leq C_3\left(\Delta t\right)^2. \tag{2.10}$$

On the other hand, from Theorem 51 it follows that if $\mu\in\mathfrak{M}_{d,M}^+[0,s_{max}]$ then

$$\rho_F(\mu,\mu\blacktriangledown)\leq\|\mu\|M^{-2}. \tag{2.11}$$

The idea of the following estimate is illustrated on Figure 2.6. Using triangle inequality and the semiflow estimate we obtain

$$\begin{aligned}&\rho_F\left(\mu_0\#_0^T, \mu_0 \bigcirc_{k=1}^{T\cdot N}\left(\bigstar_{(k-1)\Delta t}^{k\cdot\Delta t}\circ\blacktriangledown\right)\right)\leq \rho_F\left(\mu_0\#_0^T, \mu_0\bigstar_0^{\Delta t}\blacktriangledown\#_{\Delta t}^T\right)+\\ &+\rho_F\left(\mu_0\bigstar_0^{\Delta t}\blacktriangledown\#_{\Delta t}^T, \mu_0 \bigcirc_{k=1}^{T\cdot N}\left(\bigstar_{(k-1)\Delta t}^{k\cdot\Delta t}\circ\blacktriangledown\right)\right)\leq Lip(\#)\rho_F\left(\mu_0\#_0^{\Delta t}, \mu_0\bigstar_0^{\Delta t}\blacktriangledown\right)+\\ &+\rho_F\left(\left(\mu_0\bigstar_0^{\Delta t}\blacktriangledown\right)\#_{\Delta t}^T, \left(\mu_0\bigstar_0^{\Delta t}\blacktriangledown\right)\bigcirc_{k=2}^{T\cdot N}\left(\bigstar_{(k-1)\Delta t}^{k\cdot\Delta t}\circ\blacktriangledown\right)\right).\end{aligned}$$

By inequalities (2.11) and (2.10) applied to the first term we conclude that

$$\begin{aligned}\rho_F\left(\mu_0\#_0^T, \mu_0 \bigcirc_{k=1}^{T\cdot N}\left(\bigstar_{(k-1)\Delta t}^{k\cdot\Delta t}\circ\blacktriangledown\right)\right)&\leq C_4(T)\cdot(N^{-2}+M^{-2})+\\ &+\rho_F\left(\left(\mu_0\bigstar_0^{\Delta t}\blacktriangledown\right)\#_{\Delta t}^T, \left(\mu_0\bigstar_0^{\Delta t}\blacktriangledown\right)\bigcirc_{k=2}^{T\cdot N}\left(\bigstar_{(k-1)\Delta t}^{k\cdot\Delta t}\circ\blacktriangledown\right)\right).\end{aligned} \tag{2.12}$$

Notice that the upper bound consists of the term which is of order $\mathcal{O}(N^{-2}+M^{-2})$ and a term, which is equal to the error of sEBTc algorithm for a shorter time period, $T-\Delta t$. Therefore, by induction we obtain

$$\rho_F\left(\mu_0\#_0^T, \mu_0 \bigcirc_{k=1}^{T\cdot N}\left(\bigstar_{(k-1)\Delta t}^{k\cdot\Delta t}\circ\blacktriangledown\right)\right)\leq C_5(T)\left(N^{-1}+NM^{-2}\right).$$

□

Corollary 99. *sEBT and sEBTc algorithms have the same rate of convergence if $N=M$.*

Proof. If $N=M$ then $N^{-1}+NM^{-2}=\mathcal{O}(N^{-1})=\mathcal{O}(\Delta t)$. □

Numerical tests aiming at the comparison of sEBT and sEBTc in terms of accuracy and efficiency have been conducted on the following parameters:

$$\begin{aligned}g(s) &= 10\left(1-s\right),\\ m(s) &= s^2,\\ \beta(s) &= s.\end{aligned}$$

with the initial conditions equal to the Dirac mass at 0, namely $u_0=\delta_0$.

Table 2.3 presents results of the numerical analysis. The empirical order of convergence is close to 1, which confirms Theorem 98. sEBTc algorithm turns out to be significantly faster, though for given parameters N, M it induces larger error than sEBT.

Figure 2.6: Visualization of the proof of Theorem 98.

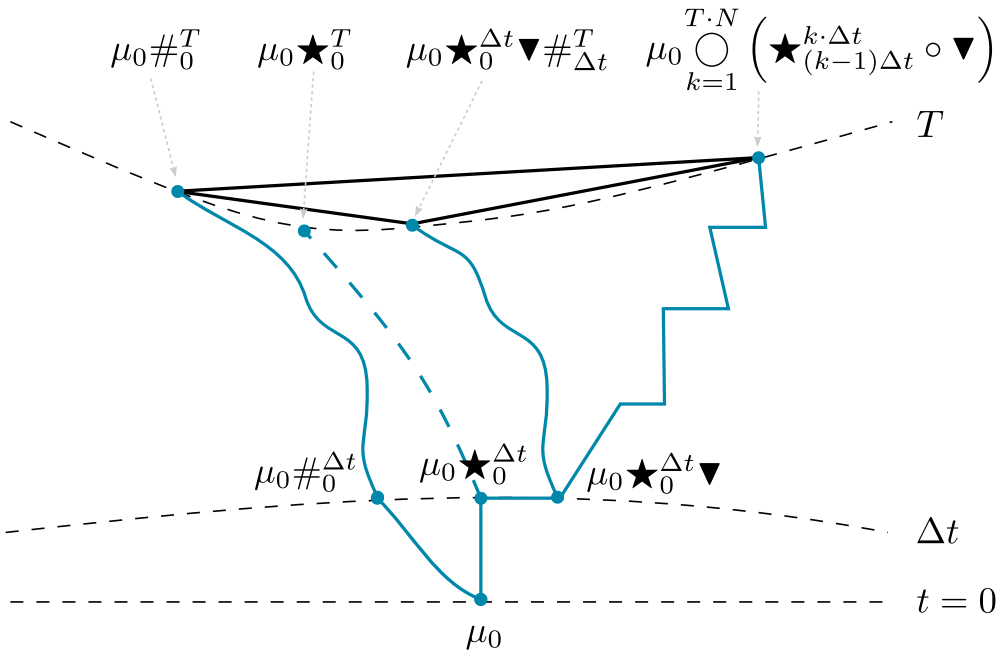


Table 2.3: Comparison of accuracy and efficiency of sEBT and sEBTc algorithms.

Parameters		sEBT			sEBTc		
N	M	error	order	CPU time	error	order	CPU time
2	8	4.74e-02		0.00s	6.02e-02		0.00s
4	16	2.60e-02	0.86	0.00s	3.19e-02	0.91	0.00s
8	32	1.35e-02	0.94	0.01s	1.64e-02	0.95	0.00s
16	64	7.10e-03	0.92	0.02s	8.65e-03	0.92	0.00s
32	128	3.64e-03	0.96	0.04s	4.51e-03	0.93	0.01s
64	256	1.83e-03	0.99	0.07s	2.27e-03	0.99	0.02s
128	512	9.05e-04	1.01	0.15s	1.13e-03	1.00	0.05s
256	1024	4.38e-04	1.04	0.30s	5.54e-04	1.02	0.09s
512	2048	2.21e-04	0.98	0.58s	2.80e-04	0.98	0.19s
1024	4096	1.14e-04	0.95	1.20s	1.45e-04	0.94	0.39s

2.2.3.3. Step functions

As shown in the Section 2.2.1 the bottleneck, in terms of accuracy, of the sEBT algorithm is the method of handling boundary conditions. By Proposition 83 birth process generates an absolutely continuous measure on the boundary, which by Proposition 69 cannot be approximated with a 1-point discrete measure with a smaller error than $\mathcal{O}\left((\Delta t)^2\right)$. The concept of this section, arising from Theorem 70, is to approximate the distribution of "young" individuals by an indicator function of the interval $[0, l^1(t)]$.

In the algorithm introduced in this section, EBT^2, the initial condition, $u_0 \in \mathfrak{M}^+[0, s_{max}]$, is approximated by a sum of a discrete measure and a step function, μ_0. Let function $m_i, n_j : [0,T] \to \mathbb{R}^{\geq 0}$ and $x_i, y_j : [0,T] \to [0, s_{max}]$ for $i \in \{1, ..., N\}$ and $j \in \{B, 1, ..., M\}$ be some function specified later, and let

$$\mu_0 = \sum_{i=1}^{N} m_i(0)\delta_{x_i(0)} + \sum_{i=1}^{M} \frac{n_i(0)}{y_i(0) - y_{i-1}(0)} \mathbb{1}_{[y_{i-1}(0), y_i(0)]}, \tag{2.13}$$

see Figure 2.7Throughout this section by $y_0(t)$ we always mean $y_B(t)$, and by $y_{-1}(t)$ we mean 0. Supports of the discrete part and the absolutely continuous part of μ_0 may overlap. Each atom of μ_0 is tracked by the system of equations analogous to (2.5)

$$\begin{cases} \frac{d}{dt}x_i(t) = g(t, \sum_{i=1}^{N} m_i(t)\delta_{x_i(t)} + \sum_{i \in I} \frac{n_i(t)}{y_i(t) - y_{i-1}(t)} \mathbb{1}_{[y_{i-1}(t), y_i(t)]})(x_i(t)) \\ \frac{d}{dt}m_i(t) = -m(t, \sum_{i=1}^{N} m_i(t)\delta_{x_i(t)} + \sum_{i \in I} \frac{n_i(t)}{y_i(t) - y_{i-1}(t)} \mathbb{1}_{[y_{i-1}(t), y_i(t)]})(x_i(t)) \cdot m_i(t) \end{cases}, \tag{2.14}$$

while each of the indicator functions is tracked by equation

$$\begin{cases} \frac{d}{dt}y_i(t) = g(t, \sum_{i=1}^{N} m_i(t)\delta_{x_i(t)} + \sum_{i \in I} \frac{n_i(t)}{y_i(t) - y_{i-1}(t)} \mathbb{1}_{[y_{i-1}(t), y_i(t)]})(y_i(t)) \\ \frac{d}{dt}n_i(t) = m(t, \sum_{i=1}^{N} m_i(t)\delta_{x_i(t)} + \sum_{i \in I} \frac{n_i(t)}{y_i(t) - y_{i-1}(t)} \mathbb{1}_{[y_{i-1}(t), y_i(t)]})(y_i(t)) \cdot n_i(t) \end{cases}. \tag{2.15}$$

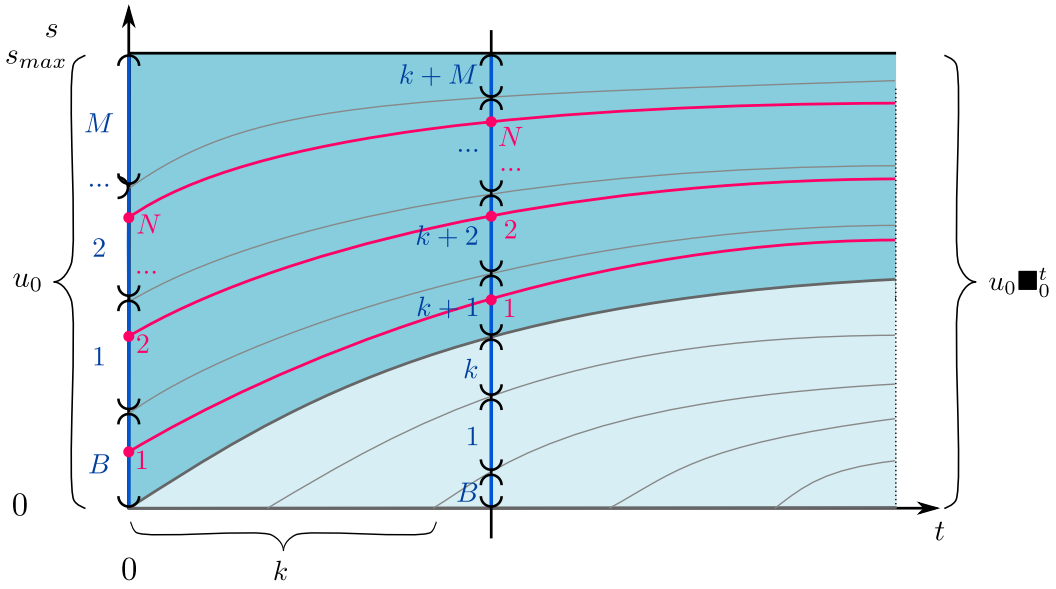
By the generalized boundary cohort, used in EBT^2, we mean indicator function $n_B(t)\mathbb{1}_{[0, y_B(t)]}$. Similarly as in sEBT, a new generalized boundary cohort is created every $\Delta t > 0$ of time, and the previous generalized boundary cohort becomes an internal cohort, tracked by (2.15). Functions $n_B(t)$ and $y_B(t)$ follow

$$\begin{cases} \frac{d}{dt}y_B(t) = g(t, \sum_{i=1}^{N} m_i(t)\delta_{x_i(t)} + \sum_{i \in I} \frac{n_i(t)}{y_i(t) - y_{i-1}(t)} \mathbb{1}_{[y_{i-1}(t), y_i(t)]})(y_B(t)) \\ \frac{d}{dt}n_B(t) = -m(t, \sum_{i=1}^{N} m_i(t)\delta_{x_i(t)} + \sum_{i \in I} \frac{n_i(t)}{y_i(t) - y_{i-1}(t)} \mathbb{1}_{[y_{i-1}(t), y_i(t)]})(y_B(t)) \cdot m_i(t) + \\ \quad + \sum_{i=1}^{N} \beta(t, \sum_{i=1}^{N} m_i(t)\delta_{x_i(t)} + \sum_{i \in I} \frac{n_i(t)}{y_i(t) - y_{i-1}(t)} \mathbb{1}_{[y_{i-1}(t), y_i(t)]})(x_i(t)) m_i(t) \\ \quad + \sum_{i \in I} \beta(t, \sum_{i=1}^{N} m_i(t)\delta_{x_i(t)} + \sum_{i \in I} \frac{n_i(t)}{y_i(t) - y_{i-1}(t)} \mathbb{1}_{[y_{i-1}(t), y_i(t)]})(y_i(t)) n_i(t) \\ y_B(k\Delta t) = n_B(k\Delta t) = 0 \end{cases}. \tag{2.16}$$

Remark 100. Existence and uniqueness of solution to the ODE system defined by equations (2.14), (2.15) and (2.16) stems from boundedness and Lipschitz continuity of parameters g, m, β upon arguments and from Lipschitz dependence of measure

$$\left(\sum_{i=1}^{N} m_i \delta_{x_i} + \sum_{i \in I} \frac{n_i}{y_i - y_{i-1}} \mathbb{1}_{[y_{i-1}, y_i]} \right) \in \mathfrak{M}^+[0, s_{max}] \tag{2.17}$$

Figure 2.7: Visualization of the EBT^2 algorithm



on m_i, x_i, n_i, y_i.

Indeed, Lipschitz dependence of (2.17) upon a single parameter x_i or y_i can be easily checked using Theorem 11. Lipschitz dependence of (2.17) upon a single parameter m_i or n_i stems directly from Proposition 7. Finally, triangle inequality for flat metric guarantees that (2.17) is also Lipschitz continuous with respect to the whole vector of parameters

$$(m_1, x_1, m_2, x_2, ..., m_N, x_N, n_1, y_1, ...)\,.$$

Similarly as in the case of sEBT, the set of indices, I, initially consists of the boundary cohort index, B, and some number, M, (possibly zero) indices of indicator functions in the initial approximations. Therefore at time $t \in [k\Delta t, (k+1)\Delta t]$ we have

$$I = \{B, 1, 2, ..., k, k+1, k+2, ..., k+M\}.$$

We shall denote the outcome of the EBT^2 algorithm at t_1 starting from initial conditions μ at t_0 by $\mu\blacksquare_{t_0}^{t_1}$.

Table 2.4 presents results of the numerical analysis conducted for the Test Case 2 described in Section 2.2.2. The empirical order of convergence of EBT^2 algorithm is close to 2, suggesting that

$$\rho_F(u_0 \#_0^T, \mu_0 \blacksquare_0^T) \leq C_1(T)\,(\Delta t)^2 + C_2(T)\rho_F(u_0, \mu_0).$$

EBT^2 induces significantly smaller error, compared to sEBT, even for large Δt since it allows to use step functions as approximation of initial conditions. In the tests presented in Table 2.4 initial condition was approximated by 8192-step function in the EBT^2 algorithm, resulting with error equal to $1.36e{-}8$, and with 8192-point measure in the sEBT algorithm, resulting with error equal to $4.33e-5$.

Table 2.4: Comparison of accuracy and empirical order of convergence of sEBT and EBT^2 algorithms.

		EBT^2	
Δt^{-1}	M	Error	Order
1	8192	$1.86e-3$	
2	8192	$4.36e-4$	2.09
4	8192	$8.98e-5$	2.29
8	8192	$2.27e-5$	1.98
16	8192	$5.13e-6$	2.14

		sEBT	
Δt^{-1}	N	Error	Order
1	8192	$1.64e-1$	
2	8192	$1.09e-1$	0.58
4	8192	$6.43e-2$	0.76
8	8192	$3.50e-2$	0.87
16	8192	$1.83e-2$	0.93

2.3. Optimal foraging model in population dynamics

In this section we apply theory described in Section 2.1 and Section 2.2 to study equation (2.1) with a specific choice of parameters reflecting growth, reproduction and mortality of *Daphnia* population under predation of a size-selective planktivorous fish in an aquatic ecosystem. It is allowed to consider a single equation for the total population without making the distinction between female and male individuals, since *Daphnia* species have a life cycle based on cyclical parthenogenesis, alternating between asexual and sexual reproduction.

In the general theory dependence of all three parameters upon time and population structure can be taken into account. Since in aquatic ecosystems where predators are present prey density levels never reach carrying capacity we shall consider a simplified model in which growth rate, g, and reproduction rate, β, are constant as functions on $[0,T]\times\mathfrak{M}^+[0,s_{max}]$ with values in $C^{0,1}[0,s_{max}]$ (independent on time and size-distribution). The argument is elaborated in Section 3.4.

It is worth mentioning the paper [33] in which an age-structure population model describing fish predation on *Daphnia* was introduced. The approach presented in this thesis allows to investigate the population in the context of arbitrary structure and not necessarily the age. In many cases, this enables to model quantities that easy to measure experimentally. In the case of *Daphnia* it is the size of an individual rather then its age that can be directly obtained from the experimental data. Moreover, the size (not age) of an individual indicates the likelihood of being detected by a forager.

A different approach to the modeling of size-structured population is described in [17], where the authors couple an ordinary differential equation for the population of roach (predators) with a McKendrick-van Foerster equation for a size-structured population of *Daphnia* (consumers), and yet another ordinary equation for algae (resources). The complex structure of this model is, however, undermined by the fact that mortality of the consumers does not take into account size-selectivity of the predator. In the model presented in this thesis predators' numerical response is neglected for the reasons discussed in detail in Section 3.1.

2.3.1. Capture rate operator

Predator-induced mortality is one of the main building blocks in the modeling of prey population dynamics. In Section 3.3 a mortality operator C_{LOW}, applicable for the case of low prey density, is derived based on the optimization of net rate of energy intake. The model of energy balance consists of:

1. the model of predator respiration rate, $R(v)$, as a function of velocity, v,
2. the model of predator post-capture acceleration costs, $A(v)$, as a function of velocity, v,
3. the model of predator reactive distance (maximum distance at which prey item can be noticed), $r(s)$, as a function of prey size,
4. the model of prey energy value, $e(s)$, as a function of prey size.

The following definition summarizes the considerations presented in detail in Section 3.3.

Definition 101. Consider a capture rate operator $C_{LOW} : \mathfrak{M}^+[0, s_{max}] \to \mathfrak{M}^+[0, s_{max}]$ defined by

$$C_{LOW}[u] = \frac{\pi v[u] r^2 u}{1 + T_h \pi v[u] \int_0^{s_{max}} r^2(\sigma) u(d\sigma)},$$

where $v : \mathfrak{M}^+[0, s_{max}] \to \mathbb{R}^{\geq 0}$ is implicitly defined as a the maximizer of $P : \mathfrak{M}^+[0, s_{max}] \times \mathbb{R}^{\geq 0} \to \mathbb{R}$

$$P(u, v) = \pi v \int_0^{s_{max}} r^2(\sigma) \left(e(\sigma) - A(v)\right) u(d\sigma) - R(v), \tag{2.18}$$

where π is the ratio of a circle's circumference to its diameter.

Acceleration cost, A, respiration rate, R, reactive distance, r, and energy value, e, are some fixed mappings of $\mathbb{R}^{\geq 0}$ to $\mathbb{R}^{\geq 0}$. In this section we make weak assumptions on the shape of these functions, which is necessary to prove well-posedness of the population dynamics equation. Concrete examples of such functions, that stem from experimental data and physical considerations, can be found in Chapter 3. Moreover, in Section 2.3.2 it is checked that these examples satisfy necessary conditions.

Condition 102. Functions A, R, r and e satisfy the following properties:

1. functions r^2 and e are Lipschitz continuous on $[0, s_{max}]$ and $A, R \in C^2(\mathbb{R}^{\geq 0}, \mathbb{R})$,
2. derivatives $A'(v)$ and $R'(v)$ are non-negative and strictly increasing,
3. $A(0) = 0$, $R(0) > 0$,
4. $\lim_{v\to\infty} A(v) = \lim_{v\to\infty} R(v) = \infty$,
5. $e(s) > 0$ and $r(s) > 0$ for $s > 0$.

We also make the following assumptions on the models of growth and birth processes:

Condition 103. Functions g, β satisfy the following properties:

1. function g is the Bertalanffy growth rate ([64]), namely $g(s) = \gamma\,(s_{max} - s)$ for some constant γ,
2. $\beta(s) = a(s - s_0)^b$ for some constants a, b, s_0 (compare [7]).

Since according to Condition 103 functions g and β do not depend on time, t, no population distribution, u, we shall often write $g(s)$ and $\beta(s)$ instead of $g(t,u)(s)$ and $\beta(t,u)(s)$ whenever Condition 103 is assumed.

2.3.2. Assumptions on parameters

Some of the assumptions in Condition 102 on functions A, R, r and e are trivially satisfied for the specific choice of parameters made in chapter 3. For example $A(v) = \frac{mv^2}{2}$, assumed in Section 3.2.6, is obviously differentiable, $A'(v) = mv$ is non-negative and increasing, $A(0) = 0$ and $\lim_{v\to\infty} \frac{mv^2}{2} = \infty$. Similarly functions $R_1 = m + qv^2$ and $R_2 = 0.003916 \cdot 10^{-0.9242+0.8494W+0.0142v+0.0189T}$ introduced in Section 3.2.1 are differentiable, satisfy $R_1(0) = m > 0$ and $R_2(0) > 0$, their limit at $v \to \infty$ is infinity and their derivatives $R_1'(v) = 2qv$, $R_2'(v) = C_1 e^{C_2+C_3 v}$ are non-negative and increasing. Energy value, $e(s) = 0.655 \cdot s^{1.56}$, introduced in Section 3.2.2 is obviously Lipschitz-continuous and positive for $s > 0$. Verifying conditions on reactive distance, r, is more complex.

Reactive distance, $r(s)$, defined in Section 3.2.3, is given implicitly by the non-negative root of the equation

$$C_1 s^2 = r^2 e^{C_2 r}$$

for some positive constants C_1 and C_2. Consequently $r(s) = 0$ if and only if $s = 0$. By implicit function theorem

$$\frac{dr}{ds} = 2C_1 s \left(2re^{C_2 r} + C_2 r^2 e^{C_2 r}\right)^{-1}$$

hence

$$0 \le \frac{dr}{ds} = \frac{C_1 s}{\left(r + \frac{C_2}{2} r^2\right) e^{C_2 r}} = \frac{C_1 s}{(\sqrt{\frac{C_1 s^2}{e^{C_2 r}}} + \frac{C_2}{2}\frac{C_1 s^2}{e^{C_2 r}})e^{C_2 r}} =$$

$$= \frac{C_1 s}{(s\sqrt{C_1} e^{\frac{C_2}{2} r} + s^2 C_1 \frac{C_2}{2})} \le \frac{C_1 s}{(s\sqrt{C_1} + s^2 \frac{C_1 C_2}{2})} \le \sqrt{C_1}.$$

We have proved that $r(s)$ is Lipschitz continuous on $\mathbb{R}^{\ge 0}$ and therefore $r^2(s)$ is Lipschitz continuous on $[0, s_{max}]$. In applications for a realistic model of reactive distance in low turbidity constant C_1 does not exceed 80.

2.3.3. Velocity functional

Since $v : \mathfrak{M}^+[0, s_{max}] \to \mathbb{R}^{\ge 0}$ models predator velocity certain regularity can be expected. In particular for small changes of prey population u predator's velocity should exhibit only small fluctuations. It is also natural to expect there exists some maximal velocity

v_{max} which cannot be exceeded. In this section we prove that $v[u]$ is correctly defined as a maximizer of $P(u,v)$ introduced in (2.18), namely we show that there exists a unique maximum of function $P(u,v)$ on $\mathbb{R}^{\geq 0}$, and that $v \in C_b^{0,1}(\mathfrak{M}^+[0,s_{max}];\mathbb{R}^{\geq 0})$. An explicit formula for v is also found for the case of $R(v)$ being a cubic function.

Theorem 104. *Under Condition 102 functional v is correctly defined and*

$$v \in C_b^{0,1}(\mathfrak{M}^+[0,s_{max}];\mathbb{R}^{\geq 0})$$

.

Proof. By simple computation we obtain

$$\frac{\partial P}{\partial v} = \pi\left\langle u, er^2\right\rangle - \pi\left\langle u, r^2\right\rangle (vA(v))' - R'(v) \tag{2.19}$$

and therefore

$$\left.\frac{\partial P}{\partial v}\right|_{v=0} = \pi\left\langle u, er^2\right\rangle - R'(0)$$

$$\lim_{v_\infty\to\infty}\left.\frac{\partial P}{\partial v}\right|_{v=v_\infty} = \lim_{v_\infty\to\infty}\left[\pi\left\langle u, er^2\right\rangle - R'(v) - \pi\left\langle u, r^2\right\rangle (A(v)+vA'(v))\right] = -\infty.$$

Since $R'(v)$, $A(v)$, $A'(v)$ are increasing functions P is concave with respect to v. Consequently, its maximum, $v[u]$, exists, is unique and always attained in the critical point or at the boundary. Moreover $v[u]=0$ if $\pi\left\langle u, er^2\right\rangle \leq R'(v)$ and $v \in (0,\infty)$ otherwise.

Let us consider function

$$F(v,\xi,\varsigma) = \xi - \zeta\,(vA(v))' - R'(v),$$

which corresponds to (2.19) with $\xi = \langle u, er^2\rangle$ and $\zeta = \langle u, r^2\rangle$. We shall prove that $v = V(\xi,\zeta)$, defined by $F(V(\xi,\zeta),\xi,\zeta) = 0$, is differentiable with respect to both arguments. By the implicit function theorem $V(\xi,\zeta)$ is differentiable on $\mathbb{R}^{\geq 0}\times\mathbb{R}^{\geq 0}$ with respect to both variables if $\frac{\partial F}{\partial v} \neq 0$ for all $v \geq 0$. Since

$$\frac{\partial F}{\partial v} = -\zeta\,(vA(v))'' - R''(v).$$

and $R''(v) > 0$ we conclude that $\frac{\partial F}{\partial v} < 0$.

Both functions r^2 and er^2 are Lipschitz continuous on $[0,s_{max}]$ and by Lemma 21 for any $u,\tilde{u} \in \mathfrak{M}^+[0,s_{max}]$

$$\rho_F(u,\tilde{u}) = \sup\left\{\langle u-\tilde{u}, f\rangle : f \in C[0,s_{max}], \|f\|_{C_b^{0,1}[0,s_{max}]} \leq 1\right\} \geq$$

$$\frac{\langle u-\tilde{u}, r^2\rangle}{\|r^2\|_{C[0,s_{max}]} + Lip(r^2)} = \frac{1}{C_1}\left\langle u-\tilde{u}, r^2\right\rangle$$

and similarly

$$\rho_F(u,\tilde{u}) \geq \frac{\langle u-\tilde{u}, er^2\rangle}{\|er^2\|_{C[0,s_{max}]} + Lip(er^2)} = \frac{1}{C_2}\left\langle u-\tilde{u}, er^2\right\rangle.$$

From the above inequalities we obtain

$$\begin{aligned} |v[u]-v[\tilde{u}]| &= \left|V\left(\left\langle u,er^2\right\rangle,\left\langle u,r^2\right\rangle\right)-V\left(\left\langle \tilde{u},er^2\right\rangle,\left\langle \tilde{u},r^2\right\rangle\right)\right| \leq \\ &\leq Lip(V)\left(\left|\left\langle u,er^2\right\rangle-\left\langle \tilde{u},er^2\right\rangle\right|+\left|\left\langle u,r^2\right\rangle-\left\langle \tilde{u},r^2\right\rangle\right|\right) \leq \\ &\leq Lip(V)\left(C_1+C_2\right)\rho_F(u,\tilde{u}). \end{aligned} \tag{2.20}$$

It is now proved that $v \in C^{0,1}(\mathfrak{M}^+[0,s_{max}];\mathbb{R}^{\geq 0})$.

To prove boundedness of v we consider

$$\frac{\partial P}{\partial v} \leq \pi\left(\|e\|_{C[0,s_{max}]}-(vA(v))'\right)\left\langle u,r^2\right\rangle - R'(v) \leq \pi\left(\|e\|_{C[0,s_{max}]}-(vA(v))'\right)\left\langle u,r^2\right\rangle.$$

Since $\frac{\partial P}{\partial v}$ is monotonously decreasing its zero is always smaller than a zero of a greater function. Therefore $v[u] \leq v_{max}$, for some constant v_{max} satisfying

$$(vA(v))'|_{v=v_{max}} = \|e\|_{C[0,s_{max}]}.$$

□

Proposition 105. *Under Condition 102 with a particular choice of*

$$R(v) = r_0 + r_1 v + r_2 v^2 + r_3 v^3$$

and $A(v) = \frac{mv^2}{2}$ *it holds that*

$$v = \begin{cases} \frac{\sqrt{4r_2^2+6(\pi\langle u,r^2\rangle m+2r_3)(\pi\langle u,er^2\rangle-r_1)}-2r_2}{3(\pi\langle u,r^2\rangle m+2r_3)} & \text{if } r_1 \leq \pi\left\langle u,er^2\right\rangle \\ 0 & \text{if } r_1 \geq \pi\left\langle u,er^2\right\rangle \end{cases}. \tag{2.21}$$

Proof. The formula follows from the fact that $v[u]$ is the root of equation

$$\frac{dP}{dv} = \pi\left\langle u,er^2\right\rangle - \frac{3\pi}{2}\left\langle u,r^2\right\rangle mv^2 - r_1 - 2r_2 v - 3r_3 v^2.$$

□

Remark 106. The condition that $R''(0) > 0$ translates to $r_2 > 0$ which guarantees that the argument of the square root in formula (2.21) is always strictly positive and hence the derivative is finite.

2.3.4. Regularity of C_{LOW}

Operator $C_{LOW} : \mathfrak{M}^+[0,s_{max}] \to \mathfrak{M}^+[0,s_{max}]$ can be viewed as a multiplier $C_{LOW}(u) = m(u)\cdot u$ defined by a given function $m : \mathfrak{M}^+[0,s_{max}] \to C^{0,1}([0,s_{max}],\mathbb{R}^{\geq 0})$. A natural question of key importance is the regularity of m.

Theorem 107. *Under Condition 102 it holds that* $m \in C_b^{0,1}\left(\mathfrak{M}^+[0,s_{max}];C^{0,1}([0,s_{max}];\mathbb{R}^{\geq 0})\right)$.

Proof. Function m can be decomposed into a functional $\mathfrak{m} : \mathfrak{M}^+[0, s_{max}] \to \mathbb{R}^{\geq 0}$ and function $r^2 \in C^{0,1}\left([0, s_{max}]; \mathbb{R}^{\geq 0}\right)$ as

$$m\left[u\right](s) = \mathfrak{m}\left[u\right] \cdot r^2(s)$$

Let $u, \tilde{u} \in \mathfrak{M}^+[0, s_{max}]$ then by a similar arguments as in 2.20 we obtain

$$\left|\mathfrak{m}\left[u\right] - \mathfrak{m}\left[\tilde{u}\right]\right| = \left|\frac{\pi v[u]}{1 + T_h \pi v[u] \int_0^{s_{max}} r^2(\sigma) u(d\sigma)} - \frac{\pi v[\tilde{u}]}{1 + T_h \pi v[\tilde{u}] \int_0^{s_{max}} r^2(\sigma) \tilde{u}(d\sigma)}\right| =$$

$$= \left|\frac{\pi\left(v[u] - v[\tilde{u}]\right) + \pi^2 T_h v[u] v[\tilde{u}] \int_0^{s_{max}} r^2(\sigma)\left(\tilde{u} - u\right)(d\sigma)}{\left(1 + T_h \pi v[u] \int_0^{s_{max}} r^2(\sigma) u(d\sigma)\right)\left(1 + T_h \pi v[\tilde{u}] \int_0^{s_{max}} r^2(\sigma) \tilde{u}(d\sigma)\right)}\right| \leq$$

$$\leq \left[\pi Lip(v) + \pi^2 T_h \left\|v\right\|_{C(\mathfrak{M}^+[0, s_{max}])}\left(\left\|r^2\right\|_{C[0, s_{max}]} + Lip(r^2)\right)\right] \rho_F(u, \tilde{u}).$$

On the other hand

$$\mathfrak{m}\left[u\right] \leq \pi \left\|v\right\|_{C(\mathfrak{M}^+[0, s_{max}])}$$

hence $\mathfrak{m} \in C_b^{0,1}(\mathfrak{M}^+[0, s_{max}]; \mathbb{R}^{\geq 0})$. It is now easy to show that

$$m \in C_b^{0,1}\left(\mathfrak{M}^+[0, s_{max}]; C^{0,1}([0, s_{max}], \mathbb{R}^{\geq 0})\right).$$

Indeed,

$$\left\|m\left[u\right] - m\left[\tilde{u}\right]\right\|_{C^{0,1}([0, s_{max}], \mathbb{R}^{\geq 0})} = \left|\mathfrak{m}\left[u\right] - \mathfrak{m}\left[\tilde{u}\right]\right| \left\|r^2\right\|_{C^{0,1}([0, s_{max}], \mathbb{R}^{\geq 0})} \leq$$

$$\leq Lip(\mathfrak{m}) \left\|r^2\right\|_{C^{0,1}([0, s_{max}], \mathbb{R}^{\geq 0})} \rho_F(u, \tilde{u})$$

and

$$\left\|m\left[u\right]\right\|_{C^{0,1}([0, s_{max}], \mathbb{R}^{\geq 0})} \leq \left\|\mathfrak{m}\right\|_{C^{0,1}(\mathfrak{M}^+[0, s_{max}], \mathbb{R}^{\geq 0})} \left\|r^2\right\|_{C^{0,1}([0, s_{max}], \mathbb{R}^{\geq 0})}.$$

□

2.3.5. Existence and uniqueness

Existence and uniqueness of weak solutions to system (2.1) under Conditions 102 and 103 stems directly from Theorem 79. Assumptions on g and β are trivially satisfied. Required regularity of m, on the other hand, results from Theorem 107.

2.3.6. Stationary state

In general, a non-trivial stationary state of (2.1) does not necessarily exist. It turns out, however, that under Conditions 102 and 103 necessary and sufficient conditions can be found. Moreover, finding the exact shape of stationary measure only requires solving two algebraic equations.

Lemma 85 provides a characterization of stationary states in the case of positive mortality of the largest individuals. Let us now suppose the contrary (lack of mortality of

the largest individuals). In a vast majority of foraging models, such as C_{LOW}, no mortality of the largest prey items ($m(t,\mu)(s_{max}) = 0$) implies no mortality of smaller items ($m(t,\mu) \equiv 0$). Consequently, by Lemma 86 null mortality in a stationary state implies null reproduction. Finally, null reproduction and a positive individual growth rate imply the lack of individuals of sizes in the range $[0, s_{max})$.

Stable existence of population consisting of individuals of a single, maximal size is not surprising under no mortality and no reproduction. In the remainder of this section we focus on the case of positive mortality and hence absolutely continuous stationary size-distributions. The density function, u, of such state satisfies

$$\begin{cases} (gu)_s = \frac{-\pi v[u] r^2 u}{1+T_h \pi v[u] \int_0^{s_{max}} r^2(\sigma) u(d\sigma)} \\ g(0)u(0) = \int_0^{s_{max}} \beta(s) u(s) ds \end{cases}$$

and therefore u can be written in the following implicit form

$$u(s) = \frac{1}{g(s)} \left(\int_0^{s_{max}} \beta(\sigma) u(d\sigma) \right) \cdot \left(e^{-\frac{\pi v[u]}{1+T_h \pi v[u] \int_0^{s_{max}} r^2(\sigma) u(d\sigma)} \int_0^s \frac{r^2(\sigma)}{g(\sigma)} d\sigma} \right). \tag{2.22}$$

Let us define

$$T_\rho(s) = \frac{1}{g(s)} \left(e^{-\rho \int_0^s \frac{r^2(\sigma)}{g(\sigma)} d\sigma} \right),$$

then clearly $u(s) = \lambda T_\rho(s)$ for some choice of $\lambda, \rho \in \mathbb{R}^{\geq 0}$.

Lemma 108. *Let g satisfy Condition 103 then $T_\rho \in L^1[0, s_{max}]$ if and only if $\rho > 0$.*

Proof. For $\rho > 0$ we obtain

$$T_\rho \leq \frac{1}{g(s)} \cdot \frac{1}{1 + \rho \int_0^s \frac{r^2(\sigma)}{g(\sigma)} d\sigma},$$

since $e^{-x} \leq \frac{1}{1+x}$ for every $x \geq 0$. Consequently, T_ρ is integrable on $[0, s']$ for every $s' < s_{max}$. On the other hand

$$\int_{s'}^{s_{max}} T_\rho(s) ds \leq \int_{s'}^{s_{max}} \frac{1}{\gamma(s_{max} - s)} \cdot \frac{1}{1 + \rho r^2(s') \int_{s'}^s \frac{d\sigma}{\gamma(s_{max} - \sigma)}} ds \leq$$

$$\leq \int_{s'}^{s_{max}} \frac{1}{\gamma(s_{max} - s)} \cdot \frac{1}{1 + \frac{\rho}{\gamma} r^2(s') \frac{s'}{s_{max} - s}} ds \leq \frac{s_{max} - s'}{\rho r^2(s') s'}.$$

□

Theorem 109. *There exists a non-trivial stationary state of equation (2.1) with $\mathbf{m} = C_{LOW}$ under Conditions 102 and 103 if and only if $\int_0^{s_{max}} \frac{\beta(s)}{g(s)} ds > 1$ and the following system of equations has a solution*

$$\begin{aligned} \int_0^{s_{max}} \beta(\sigma) T_{\rho^*}(\sigma) d\sigma &= 1 \\ \rho^* + \lambda \rho^* T_h \pi v[\lambda T_{\rho^*}(\sigma)] \int_0^{s_{max}} r^2(\sigma) T_{\rho^*}(\sigma) d\sigma &= \pi v[\lambda T_{\rho^*}(\sigma)] \end{aligned}$$

Proof. Finding the stationary state can be viewed as finding the fixed point of an operator that takes u as the argument and returns the right-hand side of equation (2.22). We have

$$\lambda = \int_0^{s_{max}} \beta(\sigma)\lambda T_\rho(\sigma)d\sigma. \tag{2.23}$$

Equation 2.23 implies that either $\lambda = 0$ (and consequently $u \equiv 0$) or

$$\int_0^{s_{max}} \frac{\beta(s)}{g(s)} \left(e^{-\rho \int_0^s \frac{r^2(\sigma)}{g(\sigma)} d\sigma} \right) ds = 1. \tag{2.24}$$

The left-hand side monotonically decreases with ρ and tends to 0 as ρ tends to infinity. Consequently, equation 2.24 uniquely defines $\rho > 0$ if and only if $\int_0^{s_{max}} \frac{\beta(s)}{g(s)} ds > 1$ (otherwise no such ρ exists, hence the only stationary state is $u \equiv 0$).

Let ρ^* satisfy 2.24. Since $u(s) = \lambda T_{\rho^*}$, the implicit formula 2.22 implies that

$$\rho^* = \frac{\pi v[u]}{1 + T_h \pi v[u] \int_0^{s_{max}} r^2(\sigma) u(d\sigma)} \tag{2.25}$$

and consequently

$$\rho^* + \lambda \rho^* T_h \pi v[\lambda T_{\rho^*}(\sigma)] \int_0^{s_{max}} r^2(\sigma) T_{\rho^*}(\sigma) d\sigma = \pi v[\lambda T_{\rho^*}(\sigma)].$$

Therefore, the conditions are indeed necessary. Conversely, it is easy to check that

$$(g\lambda T_{\rho^*})_s = -\lambda \rho^* \left(e^{-\rho^* \int_0^s \frac{r^2(\sigma)}{g(\sigma)} d\sigma} \right) \frac{r^2(s)}{g(s)} = (\lambda T_{\rho^*}) \cdot \rho^* r^2(s),$$

hence equation $(gu)_s = \mathbf{m}u$ reduces to (2.25). From the definition of ρ^* it is also clear that the second equation, namely $g(0)u(0) = \int_0^{s_{max}} \beta(s)u(s)ds$ is satisfied. □

2.4. Numerical verification of the model

In this section we investigate numerical results of McKendrick-von Foerster model with parameters satisfying Condition 102 and Condition 103, which we refer to as the model of zooplankton population.

2.4.1. Choice of parameters

Numerical results on population dynamics presented in this section are restricted to the following particular choice of parameters:

1. Mortality operator $\mathbf{m}$ is proportional C_{LOW} with: $A(v) = \frac{m_{weight} v^2}{2}$ (compare Section 3.2.6), $R(v) = r_0 + r_1 v + r_2 v^2 + r_3 v^3$ (compare Section 3.3), r given by equation (3.7) (compare Section 3.2.3), $e(s) = e_{mul} \cdot s^{e_{exp}}$ (compare Section 3.2.2),

2. Growth rate given by $g(s) = \gamma \cdot (s_{max} - s)$ for $\gamma = 0.06$ (compare [67]),

3. Birth rate given by

$$\beta(s) = \begin{cases} 0 & s < s_j \\ r_m(s - s_j)^2 & s > s_j. \end{cases}$$

Values of constants used in the simulations are presented in Table 2.5.

Table 2.5: Model parameters used in Section 2.4.3.

Parameter	Value	Unit
r_m	0.5	$\frac{ind.}{day \cdot mm^2}$
s_j	1.7	mm
s_{max}	5.2	mm
m_{weight}	12	g
T_h	1	s
I_0	10	$\frac{\mu mol}{m^2 s}$
γ	0.06	$\frac{mm}{day}$

Parameter	Value
r_0	$6.8 \cdot 10^{-3}$
r_1	$1.24 \cdot 10^{-3}$
r_2	$6 \cdot 10^{-5}$
r_3	$2.5 \cdot 10^{-5}$
e_{exp}	1.56
e_{mul}	0.655

2.4.2. Stationary state

Figure 2.8 compares theoretical result given by Theorem 109 with experimental data from [30]. The model line was computed based on the result characterizing the density of stationary state for parameters satisfying Condition 102 and Condition 103 in Section 2.3.1. Light intensity, $9\frac{\mu mol}{m^2 s}$, and predator's body length, $6 - 8cm$, were assumed to reflect the experimental setup described in [30]. Birth rate and growth rate parameters were chosen to match the species used in the experiment. Remaining parameters, including maximal prey size, birth rate and water turbidity, were fitted to the data. Evident inaccuracy in the range of small body sizes ($0.4 - 0.6$mm) and the mid-range ($0.8 - 1$mm) is likely to be caused by slower growth of the newborns and faster growth of the individuals during reproduction age, which is not taken into account in the Bertalanffy law (compare Condition 103).

The error between the size-structure measured on the 52^{nd} day of the experiment, μ_E, and the theoretically derived stationary state, μ_T, is given by

$$\begin{aligned} \rho_F(\mu_E, \mu_T) &= 0.0536, \\ \frac{\rho_F(\mu_E, \mu_T)}{\|\mu_E\|_{\mathfrak{M}[0,s_{max}]}} &= 0.0139. \end{aligned}$$

and the total number of individuals in the population, $\|\mu_E\|_{\mathfrak{M}[0,s_{max}]}$, equals $8.160\frac{ind.}{m^3}$.

2.4.3. Size-distribution dynamics

In this section numerical study of the evolution in time of the size-distribution of plankton population is conducted. Simulations were performed for parameters specified in

Figure 2.8: Comparison of stationary state density of type λT_ρ (blue line) and experimental data concerning size-distribution of *Daphnia* population subject to predation (red bars representing Dirac masses), see [30]. Plots depict the distribution of experimental *Daphnia hyalina* on the 8^{th} day (left) and 52^{nd} day (right) of the experiment.

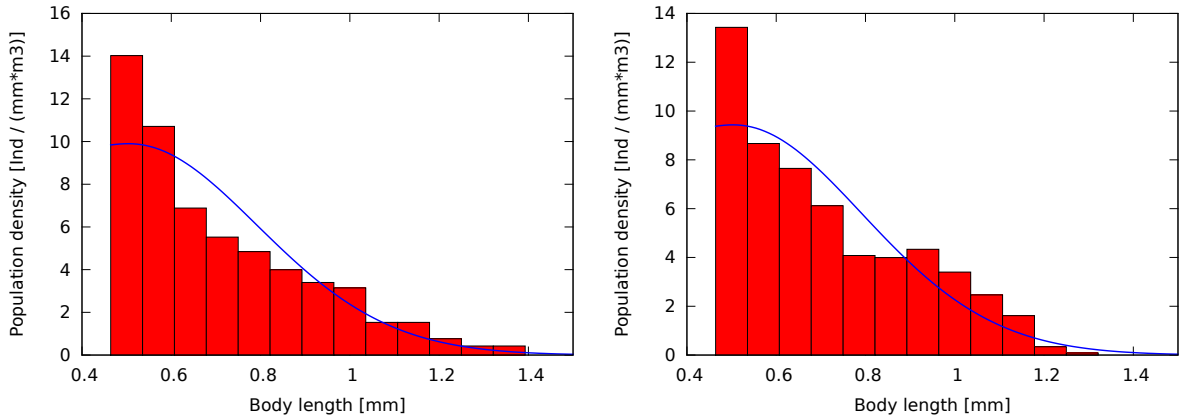


Table 2.5. Mortality was chosen to be equal $\frac{1}{10}C_{LOW}$, which reflects an average of one predator per $1m^3$ foraging during dusk and dawn, which constitute 10% of the day.

Figure 2.9 presents how the size-structure of the prey population develops in time, starting at day 1 from a single cohort of newborns. Absolutely continuous measures are depicted as plots of density functions with values on the left y-scale. Dirac deltas are shown as narrow bars whose height reflects the mass of the atom on the right y-scale. It turns out that the distribution converges to the stationary state computed using the methods from Theorem 109. Density of the stationary state is given by λT_ρ, where $\lambda = 0.12$ and $\rho = 0.31$. Figure 2.10 shows the numerical results for the same set of parameters, but starting from a uniform initial condition. The sharp peak visible at day 8 results from the birth process, which is significantly higher at the beginning, before the density of adult individuals is reduced by predation.

Figure 2.11 presents the evolution of a three-point distribution. It can clearly be seen that predator, and therefore mortality, is size-selective with high preference for larger prey items.

2.4.4. Dynamics of the total number of individuals

Numerical simulations suggest that the stationary state, characterized by Theorem 109 is not a global attractor. Figure 2.12 presents how the total number of prey individuals, namely $u(t)$ ($[0, s_{max}]$), develops in time when starting from a single cohort of newborns. It turns out that for initial densities between 0 and 9.8 individuals per dm^3 the solution converges to the non-trivial stationary state. For density equal to 0 it remains in the unstable stationary state $u = 0$, and for densities higher than $9.8\frac{ind.}{dm^3}$ it grows unlimited.

Figure 2.9: Evolution in time of the size-structure of prey population starting from a single cohort of newborns. Solution to the zooplankton population model (blue line) compared to the stationary state (brown line).

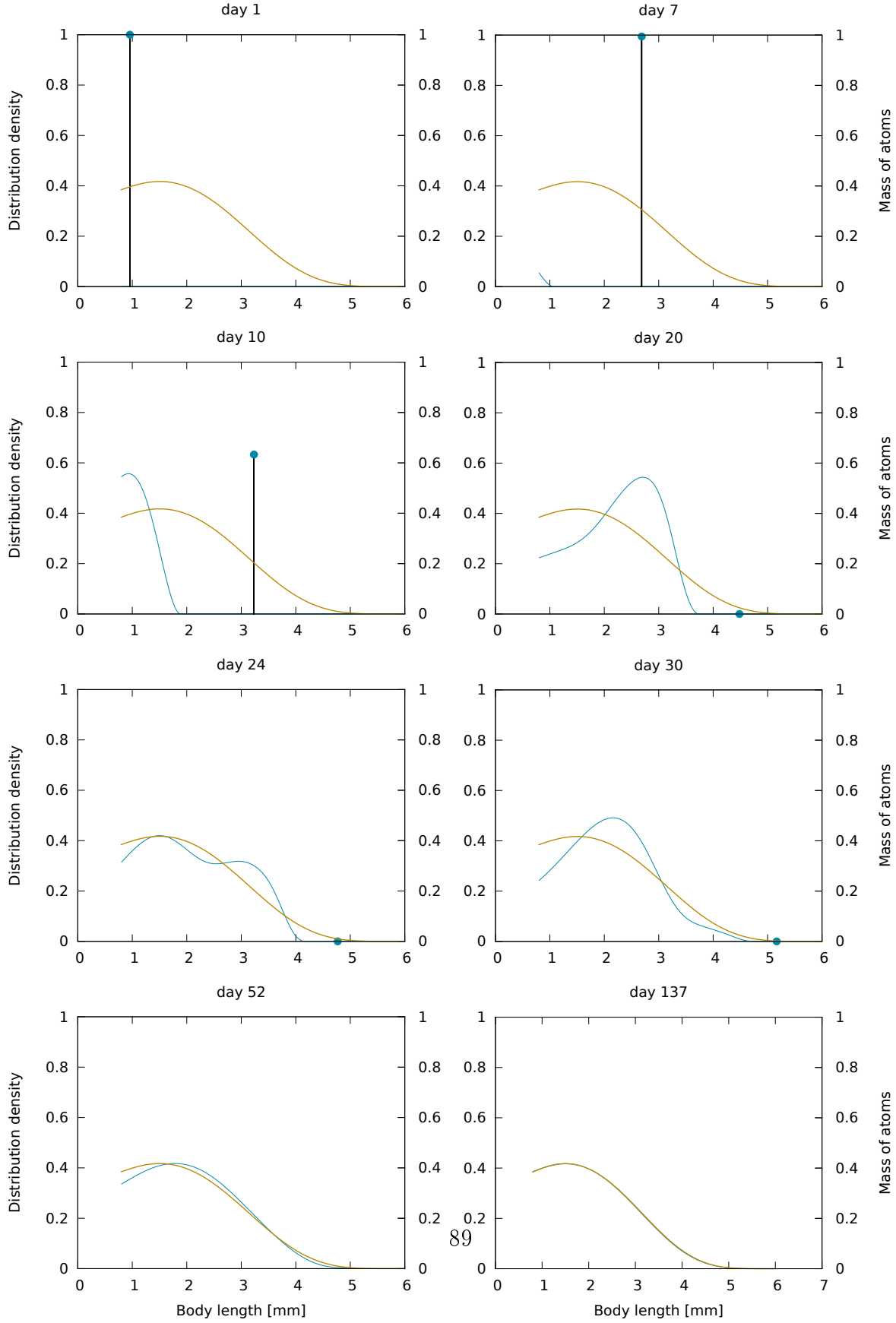


Figure 2.10: Evolution in time of the size-structure of prey population starting from a uniform distribution. Solution to the zooplankton population model (blue line) compared to the stationary state (brown line).

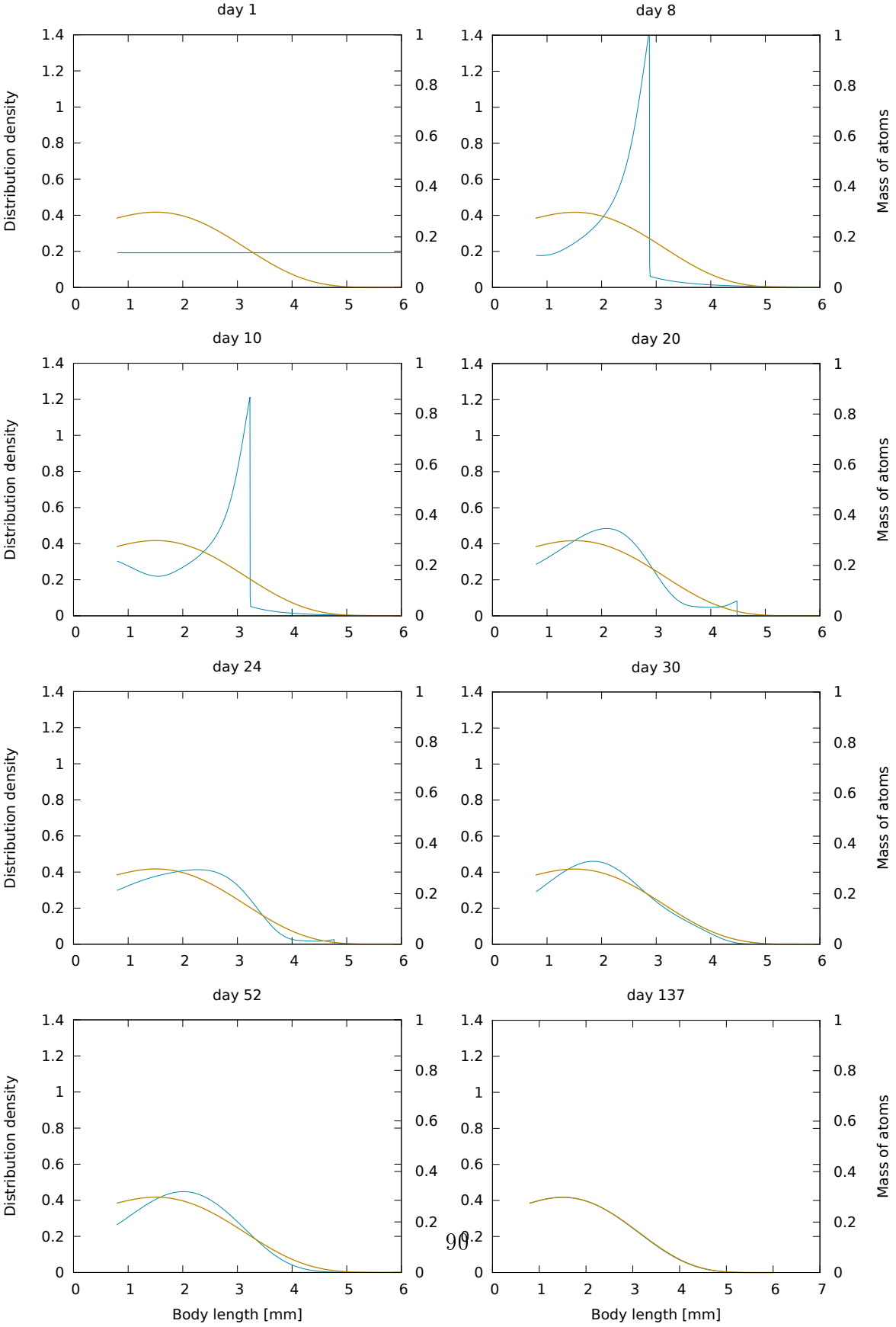


Figure 2.11: Evolution in time of the size-structure of prey population starting from a three-point distribution. Solution to the zooplankton population model (blue line) compared to the stationary state (brown line).

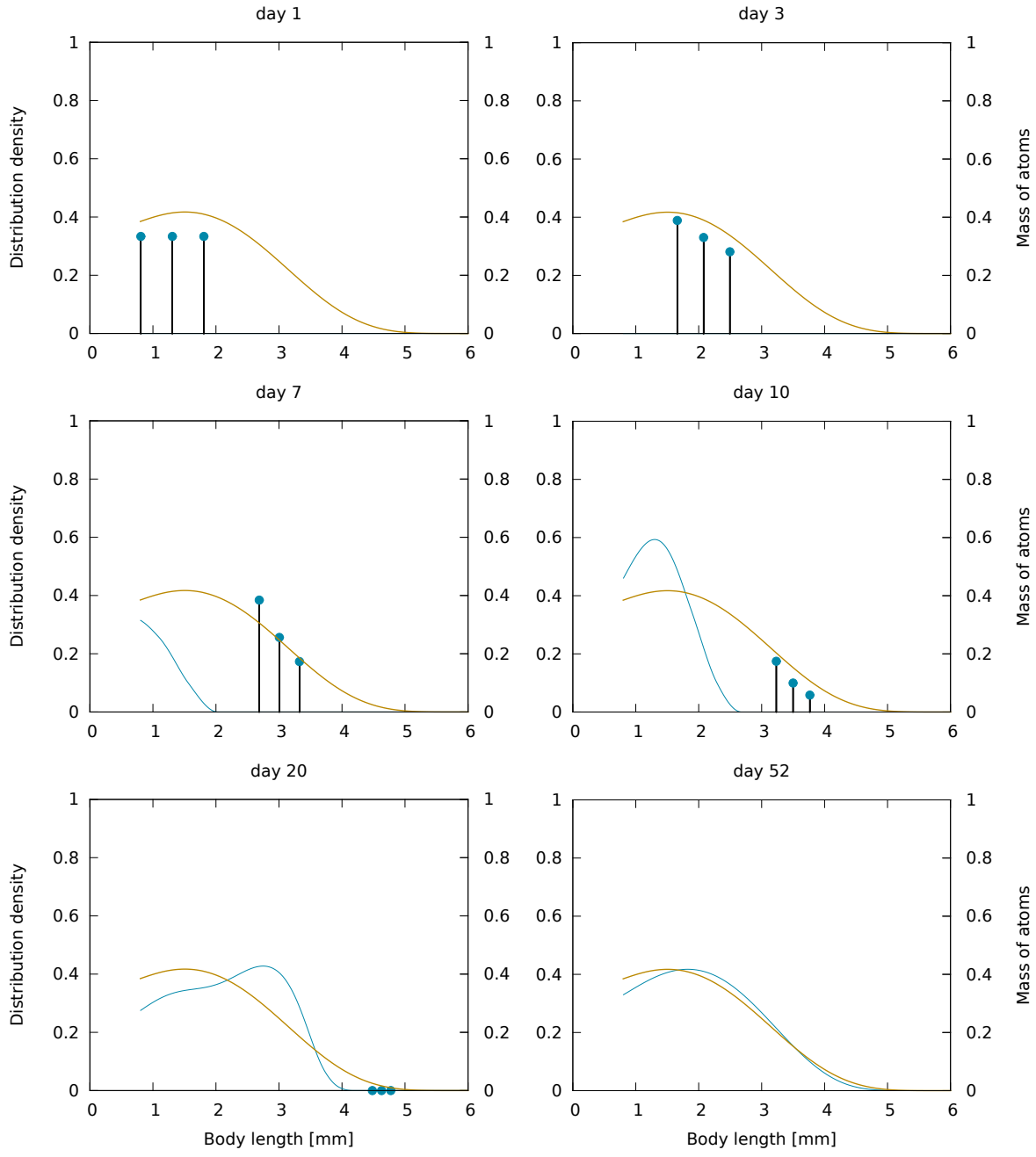
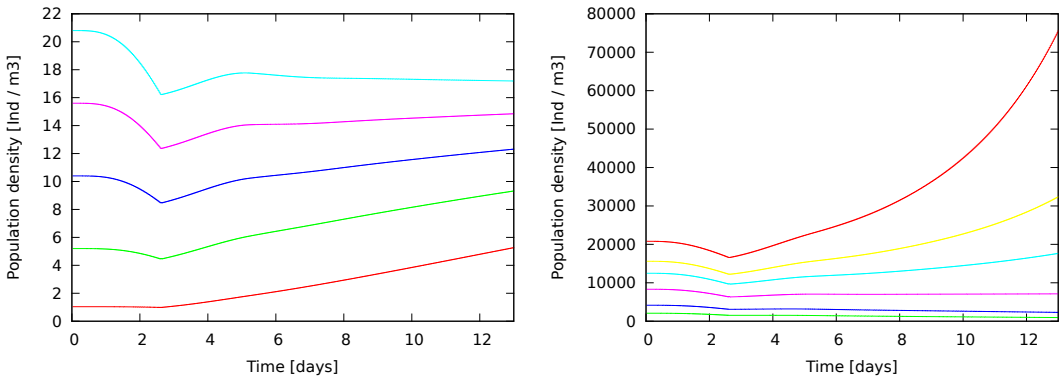


Figure 2.12: Evolution of the total number of prey individuals in the the zooplankton population model. Stable dynamics for low density initial conditions (left); instability for high densities larger (right).



Chapter 3

Foraging of a size-selective predator-harvester

The main goal of this chapter is to derive a model of predator functional response, a concept introduced by Holling in [38] to characterize different patterns of predation. The functional response is a function which assigns to the density of prey a number of prey items captured per a time unit. Some ideas developed in this dissertation were inspired by the collaboration with a team of hydrobiologists of University of Warsaw and, in particular, by the results described in [29]. Experimental evidence, obtained by the biologists, became a starting point of the study of foraging strategies.

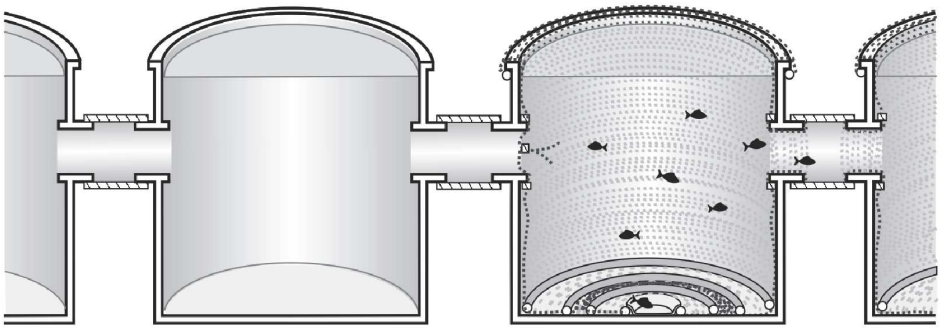
In Section 3.2 two novel simulation models of size-selective predation in ecology of freshwater ecosystems are presented. Both models are based on a bottom-up approach, in which a complex process of foraging is viewed as a composition of simpler phenomena such as predator's visual perception capability, motility and net energy balance. The models can be applied to the study of population dynamics, but are also a valuable tool for testing various hypothesis about foraging.

To give a better understanding of the matter to a reader with mathematical background Section 3.1 describes the premises on which the models were built and demonstrates the empirical data collected during author's collaboration with a team of hydrobiologists.

3.1. Experimental data

The biodiversity of an ecosystem depends on abundance of first consumers which in the case of aquatic ecosystems are mainly various species of zooplankton feeding mostly on algae. Typical species belonging to zooplankton are that of crustaceans e.g *Daphnia*. In the case of fish-free habitats where main predator feeding on zooplankton is not present the diversity of phyto- and zooplankton is more frequently attributed to resource partitioning, and resource competition . This explanation fits fish-free habitats and laboratory cultures in which a competitively-superior large-bodied *Daphnia* monopolizes resources at carrying-capacity level. However, this scenario does not match typical freshwater habitats where *Daphnia* species coexist at population densities much below those at which

Figure 3.1: The experimental system - cross-section of tanks.



resource competition would cause exclusion of competitively-inferior small-bodied taxa. This chapter is an attempt to describe quantitatively the impact of predation on the population of zooplankton.

In the experimental setup behavior of a typical freshwater planktivorous predator, 1-2 year-old roach (*Rutilus rutilus L.*) of 50-75 mm in length foraging on *Daphnia* hyalina (0.5 - 6 mm in diameter) was studied.

Prey-predator interactions between these two species are limited in scope to elimination of prey items, which allows neglecting the impact of prey population on predators. The first reason is the greatly different spatial scales of the predator and its prey. The predator, such as sardine or roach, forages kilometers each day in search of its tiny prey, while the movements of the prey are restricted to decimeters per day. The disproportion is greatest when the interactions are examined along the horizontal plane, as predation risk for a zooplankton prey depends on the light intensity and in consequence depth. The second reason is in the time scale difference due to the contrasting lifespan of the vertebrate predator and its invertebrate prey. This causes great disproportion between the reproductive numerical responses in time, which are quick in a prey population but slow in a predator population. Moreover roach and sardine individuals feed on *Daphnia* only at juvenile stage, switching to larger prey before first reproduction.

An experimental system of 4 or 8 interconnected $1m^3$ tanks, described in [29], allowed free movement of planktivorous fish between locations with different densities of *Daphnia* prey in natural mixtures of juveniles and adults (see Figure 3.1). Changes in density of *Daphnia* prey were then followed for 2-6 days. To imitate a natural field situation, fish predation was constrained by both the number of fish added to the system and how long they were allowed to feed on the *Daphnia* prey. Both parameters were adjusted to be similar to those observed in natural lake habitats where feeding by planktivorous fish is usually restricted to anti-predation windows at dusk and dawn , when the underwater light level allows them to locate their prey without being seen by piscivores.

For each feeding session, the fish were transferred to each tank in a steel bowl constituting the central part of the bottom of a cage made of nylon netting. Fish movements in one high- and one low-*Daphnia*-density tank were registered using two submerged infrared video cameras per tank, each directed at one of the two connecting windows. Analysis

Figure 3.2: Experimental data on foraging strategies published in [29]. Functional response (left) and rate of prey elimination (right). Notice the lin-log scale.

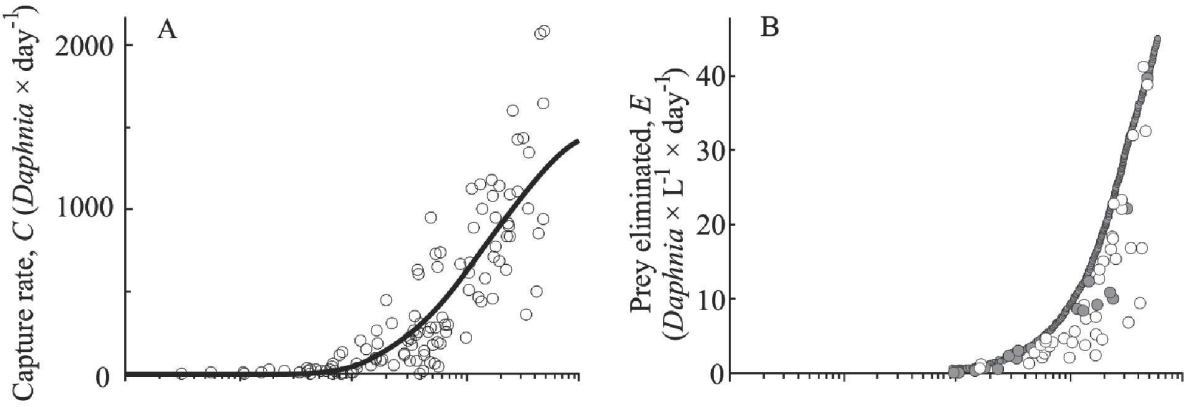
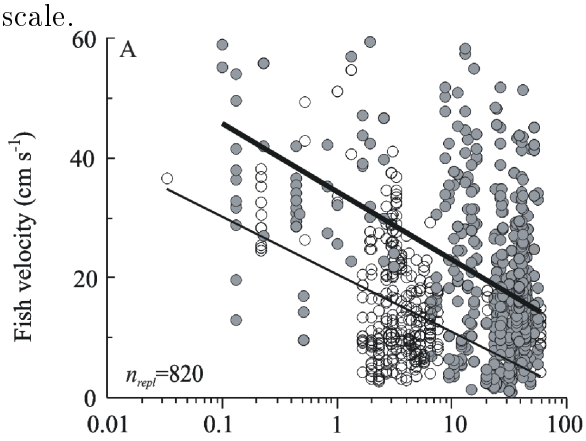


Figure 3.3: Experimental data on predator velocity published in [29]. Notice the lin-log scale.



of the resulting recordings was used to estimate fish velocity and to compute the number of fish in each of the tanks. Following the removal of fish at the end of a feeding session, the water in each experimental tank was thoroughly mixed by five upward movements of a perforated Secchi disk and samples were taken using a quantitative plankton net (6 vertical hauls removing *Daphnia* from $30l$, i.e., 3% of the tank volume) and fixed with formalin-sugar solution. *Daphnia* juveniles, adults and eggs in brood cavities were enumerated in each sample by counts made using a dissecting microscope.

Results of the experiment in a concise form of dependencies of rate of prey elimination, capture rate and predator velocity upon prey densities are presented on Figures 3.2 and 3.3. The plots do not reveal the full complexity of the results, since for instance prey elimination rate at tank A depends not only on prey density in A but also on differences in abundance between tanks. The results are, however, a starting point for a farther theoretical study.

3.2. Functional response resulting from an optimal foraging model

Since pioneering works of [51] and [21] the foraging theory was used in a vast amount of literature as a powerful tool for understanding many aspects of predator-prey interactions. It comprised the investigations of predator's optimal diet, optimal time spent foraging [1] optimal patch exploitation [51], and optimal pattern and speed of movement of a foraging predator [66]. In this theoretical study, based on the classical concept of maximization of the rate of net energy intake, we construct a model of size selective foraging. Rate of net energy intake is often used as the link between habitat use and fitness: based on the assumption that the measure of net energy intake ultimately translates into the measure of fitness (e.g., an increase or decrease in growth or reproductive output). The optimization model operating on "microecological scale" [65] describes decisions of an individual predator-harvester concerning the prey choice and speed of movement in space filled with prey items of different size and energy value. Such a framework may refer to the situation of birds (e.g. siskin or swan dive) feeding in the air on insects or a pelagic fish or an invertebrate predator feeding on zooplankton. In the case of a planktivorous fish feeding on Cladocera (*Daphnia*) prey remains relatively stationary and therefore its motility and defense during a predator's attack may be neglected. Moreover nearly all freshwater fish are plankton harvesters during the early stages of life and most remain planktivorous for a year or two before switching to either piscivory or to airborne and benthic resources. Due to an anti-predation window effect, juvenile predators rarely become satiated and as a consequence it seems to be justified to assume that optimization of the net energy intake is a fundamental factor determining individual fitness. The novelty of our approach is the optimization of the rate of net energy intake as a function of predators velocity and prey selectivity contrary to the most of earlier works in which it is assumed that encounter rate and search costs are fixed constraints independent on how quickly the predator moves. The role of choice of optimal velocity as a part of foraging strategy was argued in the case of foraging birds [37], pelagic plantktivorous fish [78] as well as in [66] in more general context. Our approach enables taking into account post-capture acceleration costs, depending indirectly on water viscosity and temperature, which seem to be a crucial constraint imposed on the predator's behavior in a low density habitat. The acceleration costs are reported as the main factor explaining differences in predator selectivity pattern in the case of small-scale homogeneous prey distribution and that of large-scale systems with heterogeneous prey distribution, as indicated recently in [53]. Our study also casts a new light on the macro ecological population level analyzing the stationary size structure of prey population.

The best known model of optimal foraging, developed by [14] and described in the monograph by [70], concerns the predator foraging on a number of prey categories whose encounter rates are given a priori as parameters. Moreover, prey items from each category have their energy values and handling times assigned. Searching for prey is assumed to cause a constant energy loss per unit of time, so despite being based on optimal foraging theory, the model does not take into account the contribution of the predator's energy expenditure due to the movement towards attacked prey items. This cost depends, in

particular, on the prey distribution in space and affects the total energy balance and optimal prey choice. In our approach a particular attention is paid to the choice of optimal velocity as a key factor contributing to predator's total energy loss. To our best knowledge the only works which account on the velocity as a crucial component of the optimal strategy are [37, 78, 18]. The latter study is based on the multi-prey functional response and similarly to Section 3.2.1 average velocity optimization is analyzed.

This section contains results described in [41]. In particular a novel, low-level, simulational model is proposed. It predicts individual forager's decision-making process including both velocity and prey choice. In this approach functional response, prey selectivity and the predator's trajectory arise from these basic decisions.

3.2.1. The case of unstructured prey population

Before introducing the main model of this part we consider a simplistic situation wherein a predator forages on an unstructured prey population. It illustrates how optimal foraging stabilizes prey-predator interactions. The classical Holling disk equation [38] can be used to derive a predator's rate of net energy intake - a quantity which can be maximized as a function of predator's velocity. Consequently the optimal predator's velocity may be expressed as a function of prey density. The optimal velocity inserted into the Holling Disk equation yields a functional response which reflects prey consumption per unit of time for an optimally foraging predator. This approach was already applied in [18] where Holling type III functional response was argued to be a consequence of optimization of predator's velocity in search. We go farther in this direction assuming a wider range of possible swimming costs and taking into account that some amount of predator's energy is spent on post-capture acceleration. Moreover we assume a more precise division of foraging time into the part devoted to searching and that devoted to prey consumption. Contrary to the aforementioned paper, we perform numerical simulations which show how the functional response is shaped depending on the particular assumptions on the cost functions.

According to the classical foraging theory, the rate at which a cruising predator encounters immobile and indistinguishable prey items is

$$\pi r^2 v N$$

where r is the reactive distance, v is the predator's velocity and $N \geq 0$ is the prey density. By the reactive distance we mean the maximum distance at which a prey item of a given size is perceivable by the predator under typical light intensity and water turbidity conditions. Assuming the handling time T_h and the attack probability a the capture rate reads

$$F(v, N) = \frac{a\pi r^2 v N}{1 + a\pi r^2 v N T_h} \tag{3.1}$$

which is known as Holling type II functional response. Notice that when $a = 1$ all prey are captured upon encounter. Then owing (3.1) and assuming the rate of velocity-dependent metabolic cost $R_i(v)$, the average energy content of prey item e and post-capture acceleration cost $A(v)$ we obtain the rate of net energy gain

$$P(v,N) = (e - A(v))\,F(v,N) - R_i(v)\gamma - R_i(0)(1-\gamma) \tag{3.2}$$

where $\gamma = 1 - T_h F(v,N)$ is a fraction of the foraging time spent on searching with velocity v while $1-\gamma$ is the remaining fraction of time, which is spent consuming prey. Formula (3.2) for the rate of net energy intake is based on the same reasoning as in [18], but the effects of stopping and accelerating are incorporated into the equation. Notice that for $T_h = 0$ formula (3.2) agrees with the simplified model introduced in (2.18) in Section 2.3.

The energy loss $R_i(v)$ denotes a basic metabolic cost and a swimming cost. In [78, 66, 77] it was proposed

$$R_1(v) = m + qv^2 \tag{3.3}$$

which is also assumed (in a slightly more general form) in [18] while [52] assume

$$R_2(v) = 0.003916 \cdot 10^{-0.9242+0.8494W+0.0142v+0.0189T} \tag{3.4}$$

where T is the temperature in Celsius, v is the velocity in meters per hour and $W = \log_{10}(0.001 \cdot w)$, where w is the body weight in kilograms. Having no experimental data on values of parameters m and q we calibrate them so that $R_1(0) = R_2(0)$ and the difference between the models is minimal. We assume the post-capture acceleration cost is equal to the physical value of the predator's kinetic energy, $\frac{wv^2}{2}$. Now we are in a position to apply the concept of optimal foraging. To this end given prey density, N, we find optimal velocity, $v_{\mathrm{opt}}(N)$, for which the rate of net energy gain, P, attains its maximum. It is easy to check analytically that such a maximum is uniquely determined for both cases (3.3) and (3.4), see Section 2.3.3. Optimal predator's functional response is obtained by setting v_{opt} in the place of v in (3.1).

The dependence of optimal velocity v_{opt} upon prey density for two different formulas describing metabolic costs of swimming, (3.4) and (3.3), was computed numerically and depicted in Figure 3.4a. In the case of (3.4) there exists a range of low prey densities where $v_{\mathrm{opt}} = 0$, which may be interpreted as the situation in which the predator chooses to stop foraging because of low light level, low prey abundance, or high water turbidity. The effect of vanishing v_{opt} also implies existence of a marginal density below which no prey items are captured. Correspondingly, the rate of net energy gain $P(v_{opt}, N)$ and the capture rate $F(v_{opt}, N)$ at depicted on Figures 3.4b and Figure 3.4c.

The stabilizing effect of prey refuges is a well known phenomenon since the experiment reported by [25] and theoretical study of the Lotka-Volterra model by [59] and [72]. Further studies of prey-predator-interaction stabilization in the context of optimal foraging were recently described in [47]. The meaning of prey refuges can be observed even in the simplest model of prey population dynamics:

$$\frac{d}{dt}N = bN - F(v_{\mathrm{opt}}, N) \cdot M$$

where b is the rate of birth coefficient and M is the number of predators. Such a model may be applied in the case when the life span of predator is much longer than that of the

prey and predators' numerical response is neglected. All of these assumption are justified in the case of planktivorous fish feeding on Cladocera [33]. Clearly, if M is larger than some critical value of predator density M_c, then there is a stable steady state at low density level $N_s > N_r$. It was checked numerically that N_s weakly changes with increase of $M > M_c$ (see Figure 3.5b). It results from the steep growth of function $F(v_{opt}, N)$ for N close to density threshold N_r from the right-hand-side (see Figure 3.4c).

We conclude that in this simplistic example the density level of prey in the steady state is mostly determined by the averaged size of prey item rather then abundance of predator c.f. [33]. It also confirms the hypothesis proposed in [28] that in the presence of planktivorous fish in lake the density levels of zooplankton are species specific and correspond to the average body size. The lower the species specific prey size the higher the corresponding threshold density level. From the results depicted in Figure 3.5a we obtained a power law $N_r \approx 0.0063 \cdot s^{-1.86}$. Figure 3.5b shows that the density of predators, unless extremely low, influences the steady state insubstantially.

Figure 3.4: Numerically computed dependence of predation characteristics upon prey density with basic metabolic cost (3.4) - (solid line)- and (3.3)- dashed line -): (a) optimal velocity - $v_{opt}(N)$, (b) rate of net energy intake $P(v_{opt}, N)$, (c) capture rate $F(v_{opt}, N)$ versus prey density (logarithmic scale).

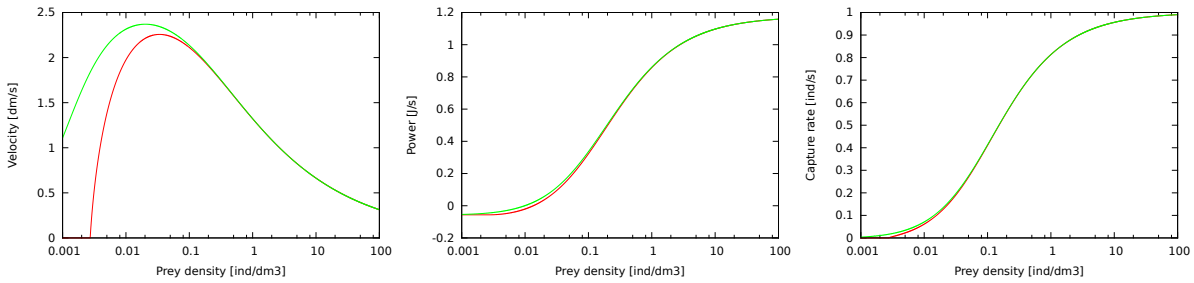
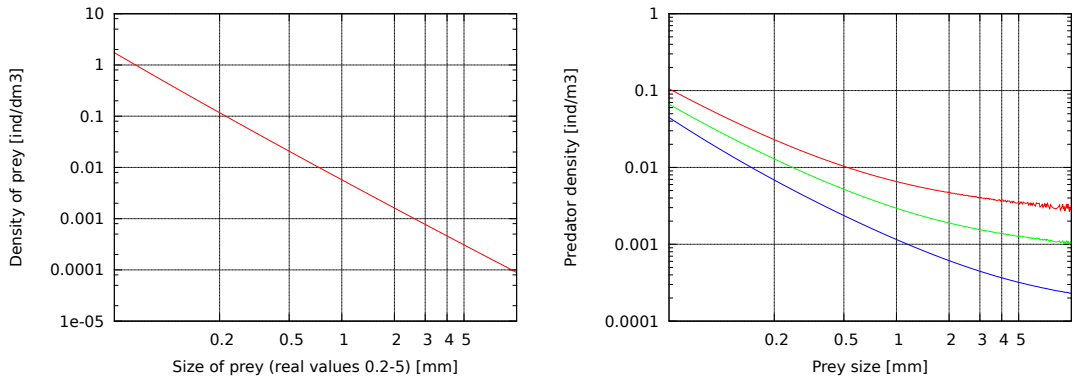


Figure 3.5: Numerically computed dependence of N_r (prey density below which predation is unprofitable) (logarithmic scale) upon prey size (a). The number of predators per $1m^3$ - M for which N_s (prey density in the steady state) is larger than N_r by 5% (red line), 50% (green line), 100% (blue line) (b). Energy of a single prey item assumed as in Section 3.2.2.



3.2.2. Energy balance of a foraging predator

In order to introduce a size structure we need to find dependence of variables such as energy intake or reactive distance upon prey size. We assume that the predator, far from being satiated, searches for prey by being in a constant motion as long as it is potentially beneficial. For each encountered prey of size s at distance $d \leq r(s)$ approached with velocity v a possible net energy intake, E, is given by

$$E = E(s, v, d) = ap(s)e(s) - A(v) - R(v)\frac{d}{v} - R(0)T_h, \tag{3.5}$$

where $e(s)$ is the energy value of prey item of size s, a is the assimilable portion of energy, $p(s)$ is the consumption success, $A(v)$ is the amount of energy needed to accelerate to velocity v just after capture, and $R(v)$ is the respiration rate when swimming with velocity v. In this paper, following [52] we assume that the energy value of a prey item (*Daphnia*) in Joules equals

$$e(s) = 0.655 \cdot s^{1.56},$$

where s is expressed in millimeters. The rate of net energy intake assigned to the prey of size s being at distance d from the predator equals

$$P(s, v, d) = \frac{E(s, v, d)}{T_h(s) + \frac{d}{v}}, \tag{3.6}$$

where $T_h(s)$ is the handling time. Note that this three-parameter function P has a different meaning than the two-parameter P defined in 3.2. For each prey in the visual field volume (VFV) an optimal velocity v_{opt}, which maximizes P, can be found. Provided realistic assumptions on R and A such maximum always exists and is unique.

The impact of consumption success rate, $p(s)$, was studied extensively in [84]. It is an important factor in the cases when prey items are either very small or have an ability to escape when under attack (e.g. copepods for planktivorous fish). In what follows we assume $p(s) = 1$, for simplicity. Such an assumption reflects the case of *Daphnia*, whose relative immotility ensures high capture success. It seems that due to the difficulties in precise parametrization of defense strategies the optimal foraging theory is expected to give good predictions in the case of immobile prey, (c.f. [71]) and in the case of predator not modulating its prey-capture behavior (c.f. [8]).

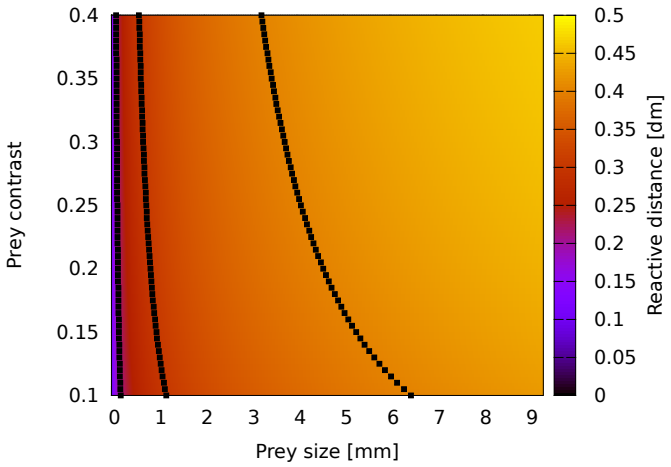
3.2.3. Reactive distance in an aquatic environment

In general, the predator's reactive distance r depends on light conditions, water turbidity, as well as features of perceivable object, in particular its contrast and size dependent cross-sectional area. It follows from theoretical considerations of ([52]) that r is the smallest number such that

$$\left(|C_0| \cdot \exp(-Cr)\right)\left(kI_0 \exp(-KZ)\right)\frac{af^2}{r^2} \geq S_t \tag{3.7}$$

where Z is the depth of foraging, K is the light extinction coefficient, C is the beam attenuation coefficient, C_0 is the inherent contrast of the prey, f is the focal length of fish

Figure 3.6: Dependence of reaction distance (color intensity) at the level of 5m under water surface upon size (x-axis) and contrast (y-axis). Curves with constant reaction distance (2cm, 3cm, 4cm) marked in black.



eye, a is the prey area, k is the ratio between radiances at retina and lens, I_0 is the light intensity under the surface and S_t is the sensitivity threshold for the detection.

The dependence of reactive distance on such factors as size, contrast of prey items, depth and turbidity implies many interesting consequences that make equation (3.7) a good starting point for many models. In this section we would like to present two examples of models addressing known ecological questions that could be built on this equation.

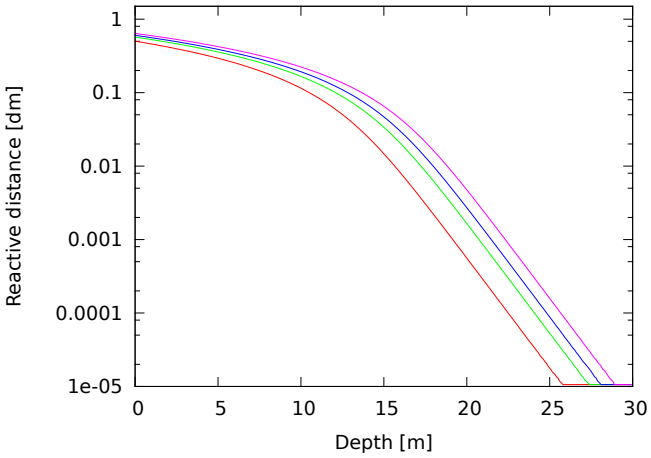
Firstly, it is worth noticing that vertical dimension plays a special role in aquatic environment because of variable light intensity. For zooplankton the layer closest to the surface is the richest in food, but also the most dangerous due to the presence of visually foraging predator. The trade-off between abundance and risk which leads to vertical distribution of copepods is studied in [27]. These considerations can be enriched by assuming realistic reactive distance model (see Figure 3.7) and abundance-dependent predator speed. Using methods described in following sections, it is also possible to take into account the size structure of copepods to obtain results on their vertical distribution based on optimal foraging theory.

Secondly, in many models variable contrast of prey items is neglected when modeling dynamics of a size-structure of a single species. It is, however, known that eggs in a brood chamber significantly increase contrast and expose individuals to a greater risk. Figure 3.6 shows the sensitivity of reaction distance with respect to the prey contrast. Using methods developed in this paper, the impact of contrast change can be assessed.

3.2.4. Expected net rate of energy intake

At low prey density when, at a given moment, there are no prey items in the predator's VFV its strategy depends on its ability of sensing prey abundance. If it's profitable to continue searching for prey then the optimal cruising speed needs to be chosen based on information about global prey distribution. We assume that the predator is capable of

Figure 3.7: Dependence of reaction distance (y-axis in logarithmic scale) upon depth (x-axis) in water of turbidity 5JTU and prey sizes: 0.5mm (red), 1.5mm (green), 2.5mm (blue), 4.5mm (pink).



assessing (or simply remembers) the overall abundance of prey in the neighborhood and chooses an appropriate optimized cruising speed. In this section we present a method of determining the optimal cruising speed that maximizes the expected value of the rate of net energy intake, P, defined in (3.6) assuming prey of size structure $u(s)$ distributed uniformly in space.

Computing the expected value of $P(\sigma, v, \delta)$ for a given v directly requires finding the joint distribution of the couple (σ, δ) of random variables, namely the size and distance to the first encountered prey. Instead of doing it in one step we use conditional expected value in order to deal with only one problem at a time.

Firstly, the distance δ to the first encountered prey of size larger than s_0 turns out to be a random variable with exponential distribution with density given by

$$g_{s_0}(\delta) = U_{s_0} e^{-U_{s_0}\delta}, \quad \delta \geq 0, \tag{3.8}$$

where

$$U_{s_0} = \pi \int_{s_0}^{s_{max}} r^2(\sigma) u(d\sigma).$$

Secondly, notice that size distribution, σ, of the encountered prey larger than s_0 is given by the probability measure $q = \frac{\pi}{U_{s_0}} r^2 u$, which is analogous to (3.12) .

Given prey size-distribution u the expected value of net energy intake (depending on predator's velocity v as well as on s_0) can be written in the general form $\mathbb{E}P(\sigma, v, \delta + r(\sigma))$, where δ is the distance the predator has to cover in order to notice the first prey larger than s_0, visible at distance $r(\sigma)$, and σ is the size of this prey. As mentioned before we use a conditional expectation to compute this value:

$$\begin{aligned}\mathcal{E}(v,s_0) &= \mathbb{E}P(\sigma,v,\delta+r(\sigma)) = \mathbb{E}\left(\mathbb{E}P(\sigma,v,\delta+r(\sigma))|\delta\right) = \\ &= \mathbb{E}\left(\int_{s_0}^{s_{max}} \frac{\pi}{U_{s_0}} r(\sigma)^2 P(\sigma,v,\delta+r(\sigma))\cdot u(d\sigma)\right) = \\ &= \pi\int_0^{\infty}\int_{s_0}^{s_{max}} e^{-U_{s_0}\delta} r(\sigma)^2 P(\sigma,v,\delta+r(\sigma))\cdot u(d\sigma)d\delta. \qquad (3.9)\end{aligned}$$

The parameter s_0 indicates a possible smallest size of the prey that could be captured. Taking the supremum over $s_0 \in [0, s_{max}]$ from 3.9represents selection of optimal marginal prey size. The optimal cruising speed is the argument for which 3.9 is maximal. Finally, our optimization procedure leads to the optimal couple

$$(v_{cruis}, s_{min}) = \max\arg_{(v,s_0)\in[0,\infty]\times[0,s_{max}]} \mathcal{E}(v,s_0).$$

In our simulation we introduce an equidistant grid on $[0, s_{max}]$ and compute v_{cruis} as a maximizer of $\mathcal{E}(v,s_0)$ for each value of s_0 of the mesh using golden section search on $v \in [0,V]$, where the upper limit V is chosen heuristically.

Note that the cruising speed computed in this section is a different notion than the optimal speed computed in Section 3.2.1 and introduced in [18]. Indeed, v_{cruis} guarantees the best expected net energy gain whenever no prey items are in VFV and therefore it should be acquired in the searching strategy. On the other hand the optimal speed obtained in Section 3.2.1 is defined as the most profitable mean velocity in a very rough averaged Holling-type model.

The cruising speed in the average model defined as the most profitable velocity maximizing the value of $P(s,v,\mathbb{E}\delta)$ coincides , as we argue below, with the optimal speed defined in Section 3.2.1 in the case of low encounter rate and unstructured population. Nonetheless, our intention is taking into account predator's behavior when no prey items are visible, rather than the behavior when the next item is precisely at an average distance.

In this paragraph we shall write r, u and e instead of $r(s)$, $u(s)$, $e(s)$ as we only consider unstructured populations. Let us put $d = \mathbb{E}\delta = \frac{1}{\pi r^2 u}$ (notice that unit of u is m^{-3}) to the rate of net energy intake $P(s,v,d)$ obtained in (3.6). For low encounter rate when it is allowed to assume that T_h is negligible compared to δ/v the rate of net energy intake $P(v,u)$ as computed in (3.2) is approximately equal to that of (3.6). Indeed in this case it follows from (3.8) that

$$\begin{aligned}P(s,v,\mathbb{E}\delta) &= \frac{e - A(v) - R_i(v)\frac{1}{\pi r^2 uv} - R_i(0)T_h}{T_h + \frac{1}{\pi r^2 uv}} \\ &\approx \pi r^2 uv\left(e - A(v) - R_i(v)\frac{1}{\pi r^2 u\cdot v} - R_i(0)T_h\right) = L\end{aligned}$$

and, on the other hand, using (3.2) and assuming $a = 1$ we find

$$
\begin{aligned}
P(v,u) &= (e - A(v))\,\frac{a\pi v r^2 u}{1 + a\pi v r^2 u \cdot T_h} - R_i(v)(1 - T_h \frac{a\pi v r^2 u}{1 + a\pi v r^2 u T_h}) + \\
&\quad - R_i(0) T_h \frac{a\pi v r^2 u}{1 + a\pi v r^2 u T_h} \approx \\
&\approx \pi r^2 u v \left(e - A(v) - R_i(v) \left(\frac{1}{\pi r^2 u v} - T_h \right) - R_i(0) T_h \right) \approx L.
\end{aligned}
$$

Finally, these considerations lead to the conclusion that when neglecting T_h we have $P(v,u) = P(s,v,\mathbb{E}\delta)$.

Both presented approaches lead to the same net rates of energy intake in the limit (with prey density tending to 0) and consequently to the same optimal velocity. The argument presented for the case of unstructured population can be generalized and it can be shown that $P(v,u) = \mathbb{E}P(s,v,\mathbb{E}\delta)$ if $T_h = 0$ also when u is a structured population and $P(v,u)$ is given by 2.18. In the next chapter, an individual based, mechanistic model of predation on a structured population is introduced. The notions of optimal cruising speed and expected net energy intake are used to model the predator's decision process.

3.2.5. Individual based model

In this section we introduce an optimal foraging model with two variants. In both of them the predator patrols a 3D environment continuously seeking for prey. When some prey items appear in VFV it then assesses the distance to each of them, and optimal velocity at which the prey item may be reached maximizing the rate of net energy intake (see Section 3.2.2). Finally the predator chooses the prey item which ensures the highest rate of the net energy intake. The appearance of prey in VFV depends on the position of predator and on the reactive distance attributed to a given prey. The geometry of VFV is taken to be a half ball around the predator's head of radius equal to the reactive distance.

In the Basic Optimal Foraging Model (BOFM), the predator's choice of particular prey item and attack velocity are based both on the information from VFV and an assessed global prey density and corresponding expected rate of net energy intake (see Section 3.2.4). Whenever there is at least one individual in VFV, either the most profitable of them is chosen for the attack or (based on global information) all of them are ignored and the more profitable ones are sought outside the VFV. We also introduce a modification of this model MOFM (long-term Memory-driven Optimal Foraging Model) which applies for heterogeneous patchy prey distribution in space. In this version the predator exhibits a transient behavior moving to a more profitable region (in terms of higher food level). In such a case, capturing prey can be considered as a side-effect and the predator decides to stop and capture a prey only if the gain compensates the additional time spent in the transient region with the reduced availability of food. Therefore we introduce the notion of anticipated energy gain, which is an energy equivalent of all the profits resulting from finding a desired place. It can be used to evaluate the loss caused by prolonging the search in region with relatively low food availability, cf. [29].

Results of simulations depicted on Figure 3.8 exhibit that the range of area patrolled by the predator as well as its average velocity increase significantly with the decrease in prey density.

Basic model for the case of homogeneous prey distribution The model of predator's behavior can be described in one sentence: the predator selects a prey (from all visible prey items) which gives the highest rate of net energy uptake. A simulation algorithm for the case of homogeneous prey distribution can be decomposed to the following steps:

1. perceive all prey items that are in predator's VFV and are larger than s_{min} (see Section 3.2.4),
2. for each prey item, individually find optimal velocity v_{opt} and compute maximal rate of net energy gain $P(s, v_{opt}, \delta)$ using (3.6),
3. choose such a prey item from VFV that guarantees maximal rate of net energy gain P,
4. move the predator to the prey with velocity v_{opt} and attack the prey,
5. keep moving the predator with velocity v_{cruis} until a point where at least one prey appears in the VFV.

In the case of lack of prey items in the VFV patrolling can still be profitable (provided $\mathbb{E}P(\sigma, v_{cruis}, \delta + r(\sigma)) > R_i(0)$). In such a case step 6 should be executed (the predator should choose to search for prey with cruising speed v_{cruis}). Otherwise, the predator may decide to rest or to continue to forage due to different reasons than instantaneous energy intake (e.g. moving to a more profitable area).

Long-term memory-driven foraging model for heterogeneous prey distribution In BOFM predator makes use of the 'knowledge' about global prey density in order to choose the optimal speed when ignoring all prey items in the VFV is profitable. In MOFM we assume, that the predator's motivation to keep moving doesn't result from the need to forage in the current location, but that there is an external reason pushing the predator to motion. An example of such scenario is a patchy environment, where the main motivation for predator's movement in low prey density comes from the need to relocate in order to find a food patch.

This variant of the model takes two additional parameters: anticipated energy gain in the searched habitat P_a and cruising speed v_a, which represent information about the heterogeneous environment available to the predator. The foraging algorithm is modified so that these two parameters are used instead of $\mathbb{E}P(\sigma, v_{cruis}, \delta + r(\sigma))$ and v_{cruis}. Namely, we obtain the modified model by substituting v_{cruis} by v_a in all steps of Section 3.2.5, as those two parameters play exactly the same role, and changing the comparison in step 1 to the following:

$$P(s, 0, 0) - \frac{A(v_a)}{T_h} \text{ against } P_a$$

Figure 3.8: Predator's trajectory (presented as a 2D projection) according to BOFM in prey density of a) 3 ind/l, b) 0.05 ind/l, c) 0.01 ind/l during 3 hours of constant foraging. For comparison of spacial scales the trajectory of high density a) is also contained in a small rectangle in left-bottom corner of b) and analogously trajectory b) is rescaled to fit in c).

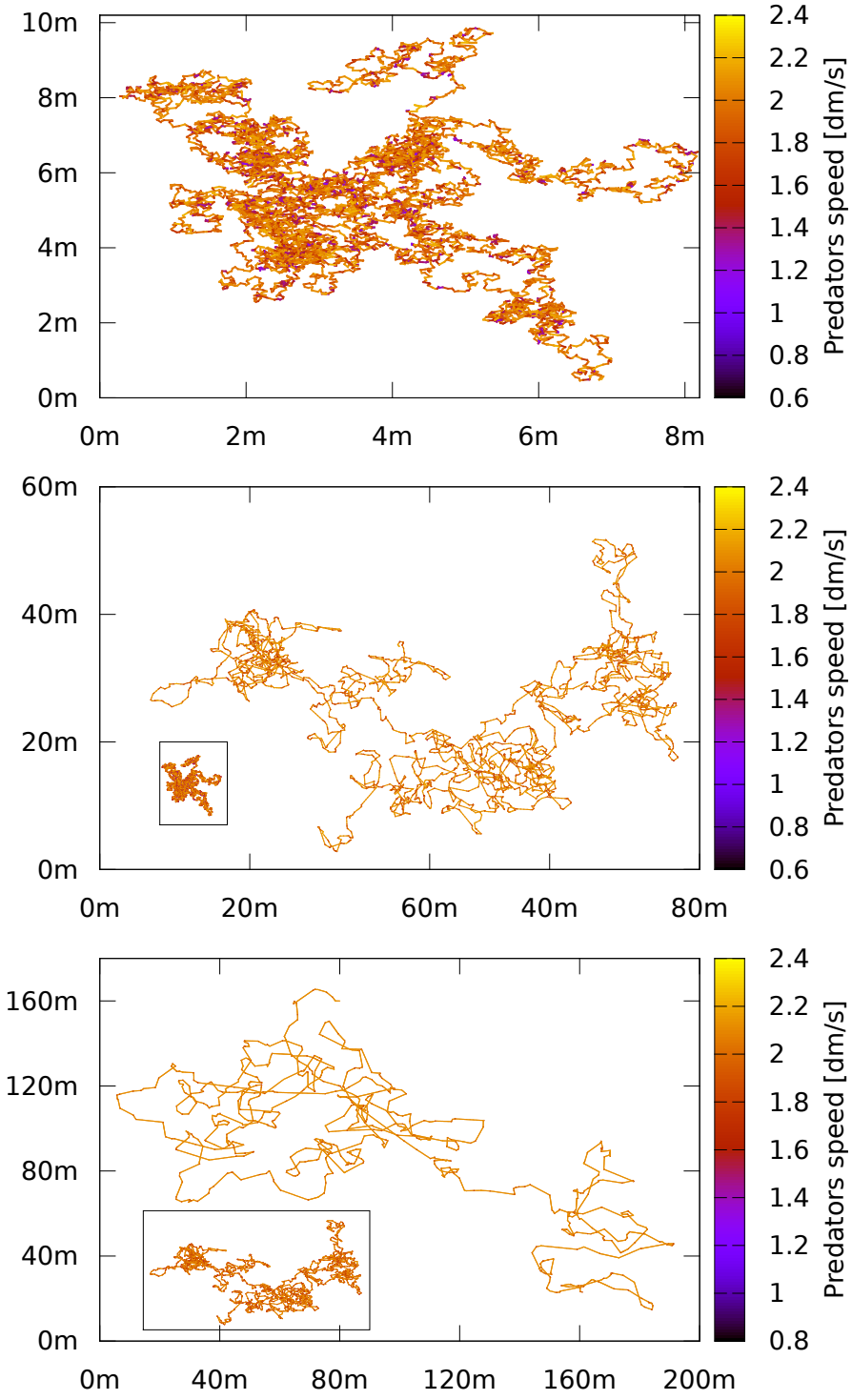
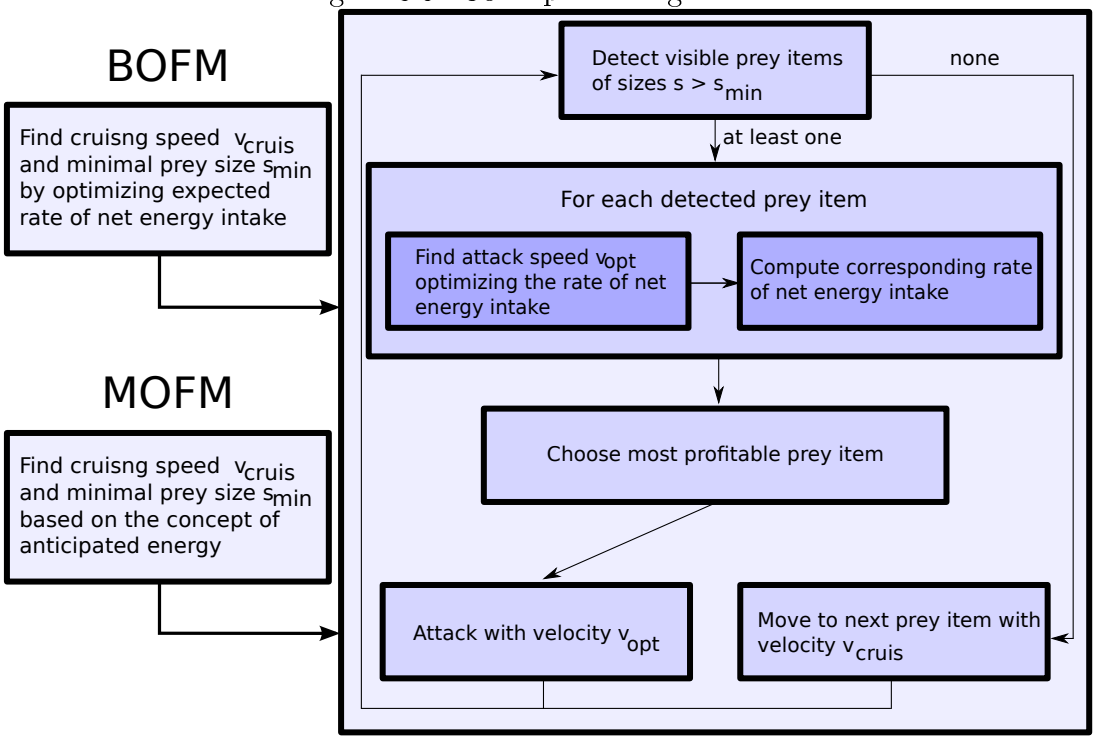


Figure 3.9: Conceptual diagram of the model



The latter formulation represents the choice between capturing a given prey item, which delays the arrival to a patch and prolongs the search, or ignoring it, which doesn't yield instantaneous gain but increases the expected profit of patch exploitation.

The main reason for introducing MOFM is to evaluate realistically and distinguish between the predator's selectivity in homogeneous and in heterogeneous prey distribution. The comparison is postponed to Section 3.2.9.2.

3.2.6. Post-acceleration costs

The energy costs due to the post-capture acceleration seem to be an underestimated factor in forager's energy budget. In fact post-capture acceleration costs have recently been taken into account in [29], where the impact of the aggregational response of predators on shaping the space distribution of prey population has been studied. Our model confirms that neglecting the acceleration costs leads to unrealistic predictions of predator's optimal velocity. Even in the simple case of optimizing foraging on unstructured population, considered in Section 3.2.1, assuming $A(v) = 0$ causes the predictions of the predator's velocity to be an order of magnitude higher then in the case of experimental data. As we were unable to find credible empirical data on the post-acceleration costs, we decided to neglect inefficiency in predator's movement and assume physically simplistic model in which energy cost of acceleration is equal to the difference of predator's kinetic energy:

$$A(v) = \frac{wv^2}{2}$$

Table 3.1: Optimal cruising speed (v_{cruis}) and expected rate of net energy gain ($\mathbb{E}P$) for varying post-capture acceleration costs

$A(v)$	$N = 0.01$	$N = 0.1$	$N = 1.$	$N = 10.$	$N = 100$
$wv^2/2$	$v_0 = 0$ $\mathbb{E}P = 0$	$v_0 = 3.13$ $\mathbb{E}P = 0.07$	$v_0 = 2.77$ $\mathbb{E}P = 0.23$	$v_0 = 2.51$ $\mathbb{E}P = 0.31$	$v_0 = 2.46$ $\mathbb{E}P = 0.34$
$wv^2/2 + 30\%$	$v_0 = 0$ $\mathbb{E}P = 0$	$v_0 = 2.79$ $\mathbb{E}P = 0.05$	$v_0 = 2.49$ $\mathbb{E}P = 0.20$	$v_0 = 2.27$ $\mathbb{E}P = 0.28$	$v_0 = 2.22$ $\mathbb{E}P = 0.31$

By this assumption we also neglect any hydrodynamical effects that may influence the cost. In fact we expect the value of $A(v)$ to depend on water viscosity and, in consequence, on its temperature. To assess possible impact of temperature on predator's strategy we investigated the dependence of optimal cruising speed and the rate of net energy gain upon post-capture acceleration costs.

The results of simulations show that the increase of post-acceleration costs by 30% yield at most 5% decline of the optimal cruising speed; compare table 3.1. The choice of 30% difference in acceleration costs presented in table 3.1 is arbitrary, but we find it relevant as the upper bound for the influence of water temperature ranging from 12^oC to 23^oC.

3.2.7. Variable handling time

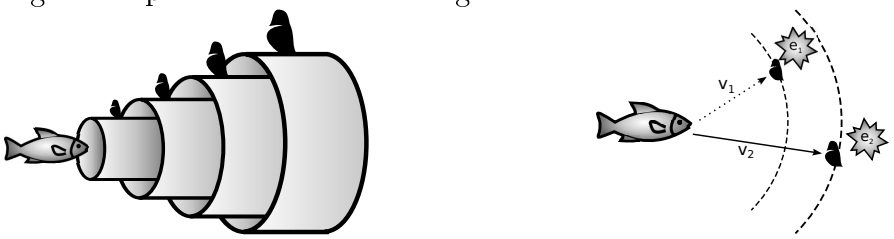
In the Holling model variable handling time depending on the prey size and prey category $T_h(s)$ is often considered. In our model the need to introduce this dependence is considerably reduced, because contrary to the case of Holling model, T_h only consists of the time necessary to capture (attack) the prey. The time needed for the predator to approach a chosen prey always depends on the size and category of the item (reactive distance depends on prey size and contrast) but is not a component of T_h in our model. On the contrary, handling time in Holling model consists of both: time needed for the predator to approach a spotted prey item and the time to capture it.

For the aforementioned reasons and the fact that body size of zooplankton is significantly smaller than predator's snout we decided to assume a constant handling time in the remaining part of the article. Described methods, however, are general enough to allow for size-dependent handling time.

3.2.8. Prey selectivity in structured population

In this section two types of selectivity of the predator are investigated: passive, resulting from the immanent selectivity of predator's sight; and active, resulting from predator's choice as depicted in Figure 3.10. These two modes of feeding refer to that distinguished in the literature as the reactive-field-volume model and the apparent size model respectively, [81, 20, 83]. To achieve this we use Jacobs selectivity index and also introduce its modification in order to investigate active and passive selectivity independently. We also checked that the Manly index yields qualitatively same results.

Figure 3.10: Passive selectivity resulting from vision limitations on the left, active selectivity resulting from optimal choice on the right



As a tool to measure total selectivity we use: Jacobs selectivity index, [42], defined as

$$D_i = \frac{r_i - p_i}{r_i(1-p_i) + p_i(1-r_i)} \tag{3.10}$$

where r_i is the probability that randomly chosen prey item selected by the predator is in i^{th} category, and p_i is the probability that randomly chosen prey item from the environment is in i^{th} category. Probabilities r_i and p_i may be approximated by empirical proportions.

As a measure of active selectivity we introduce Jacobs active selectivity index, defined as

$$D'_i = \frac{r_i - q_i}{r_i(1-q_i) + q_i(1-r_i)} \tag{3.11}$$

where q_i is the probability that a randomly chosen encountered prey item is in i^{th} category.

3.2.8.1. Passive selectivity

Following [20] by passive selectivity, we mean the phenomenon of encounter rate being prey size-dependent, and prey being captured at the rate proportional to the encounter. This phenomenon can be fully described by simple formulas derived in this section.

The encounter rate of prey items of size between s and s' in a structured population with a given distribution $u(s)$ can be written as $\pi v \int_s^{s'} r^2(\sigma)u(d\sigma)$. Therefore probability distribution of the encountered prey sizes, and consequently size distribution of captured prey is given by the normalization of this value, namely

$$q_u(E) = \frac{\int_E r(s)^2 u(ds)}{\int_0^{s_{max}} r(s)^2 u(ds)}. \tag{3.12}$$

Similarly, the size distribution of a randomly chosen prey item in the environment is given by

$$p_u(E) = \frac{u(E)}{u[0, s_{max}]}.$$

Jacobs index D_i of passive selectivity may be computed by inserting $p_i = u(\Omega_i)/u[0, s_{max}]$ and $r_i = \int_{\Omega_i} r(s)^2 u(ds) / \int_0^{s_{max}} r(s)^2 u(ds)$ into (3.10) where Ω_i is a range of sizes which belong to the investigated category.

3.2.8.2. Active selectivity in the case of low encounter rate

In low prey abundance or high turbidity, the optimal foraging model becomes much simpler, as the number of prey items in the predators visual volume is larger than 1, with only a very small probability. This is the case frequently met in turbid pools [26]. Under these circumstances the predator selects its victim actively only in the sense that it can ignore a certain prey item.

Holling functional response was originally formulated for a single prey type and may be extended to the case of arbitrarily many prey categories [20, 3]. Assuming common handling time for all prey items the capture rate of prey of type i reads

$$\frac{\alpha_i E_i}{1 + T_h \sum_i \alpha_i E_i}, \tag{3.13}$$

where E_i is the encounter rate of prey of type i and α_i is the attack probability upon encounter. This result can be reconsidered in the framework of measure theory as a definition of a capture rate operator, $C : \mathfrak{M}^+ \to \mathfrak{M}^+$, characterizing predation. Such an operator takes a population size-distribution as an arguments and returns a size-distribution of eliminated items in a time unit. Formula (3.13) can be rewritten as

$$C\left[\sum_i m_i \delta_{s_i}\right] = \frac{\sum_i \alpha_i E(s_i, m_i)\delta_{s_i}}{1 + T_h \sum_i \alpha_i E(s_i, m_i)},$$

where $E(s_i, m_i)$ is the encounter rate of prey of size s_i whose density in the environment is equal m_i. This formula can be generalized to any input measure $u \in \mathfrak{M}^+$

$$C[u] = \frac{\pi v \alpha r^2 u}{1 + T_h \pi v \int_0^{s_{max}} \alpha(\sigma) r^2(\sigma) u(d\sigma)}. \tag{3.14}$$

for a given piecewise continuous attack probability function $\alpha : [0, s_{max}] \to [0, 1]$. If u is absolutely continuous with respect to Lebesgue measure then the density of $C(u)$ is given by

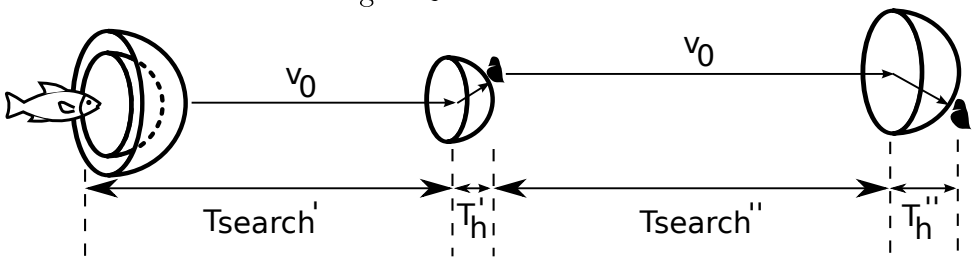
$$\frac{dC}{d\mathcal{L}} = \frac{\pi v \alpha(\sigma) r^2(\sigma) \frac{du}{d\mathcal{L}}(\sigma)}{1 + T_h \pi v \int_0^{s_{max}} \alpha(\sigma) r^2(\sigma) u(\sigma) d\sigma}.$$

In the models introduced in this paper, the fish attacks encountered prey of size s if and only if $s \geq s_{min}$ in the case of BOFM or $P(s, 0, 0) - \frac{A(v_a)}{T_h} \geq P_a$ in the case of MOFM. The simulation of BOFM is therefore expected to give very similar results as Holing-type model (3.14) with $v = v_{cruis}$ and

$$\alpha(s) = \begin{cases} 1, & s \geq s_{min}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.15}$$

for low encounter rates. Notice that in this case parameter $\alpha : \mathfrak{M}^+[0, s_{max}] \times [0, s_{max}] \to [0, 1]$ implicitly depends on prey density and its size structure u. Analogously, MOFM is expected to yield similar results as the Holling-type model with $v = v_a$ and

Figure 3.11: Low encounter rate scenario



$$\alpha(s) = \begin{cases} 1, & P(s,0,0) - \frac{A(v_a)}{T_h} \geq P_a, \\ 0, & \text{otherwise.} \end{cases} \tag{3.16}$$

For the comparison we refer to Figure 3.16.

In Holling-type models T_h includes both the time necessary for the attack and the time needed to swim through the reactive distance (compare Figure 3.11), while in BOFM and MOFM T_h only consists of the attack time. This difference can be mitigated by introducing size-dependent handling time to the Holling model.

Note that in the case of low encounter rate, both BOFM and MOFM obey the classical Zero-One Rule [70], which states that a type of prey is either always taken upon encounter or never taken upon encounter.

3.2.8.3. Active selectivity in the case of high encounter rate

Active selectivity becomes a much more complex phenomenon when high encounter rate occurs. Identically as in the case of low encounter rate, prey items smaller than some critical value are never attacked. Large enough prey, on the other hand, are only attacked if there is nothing even more profitable in VFV.

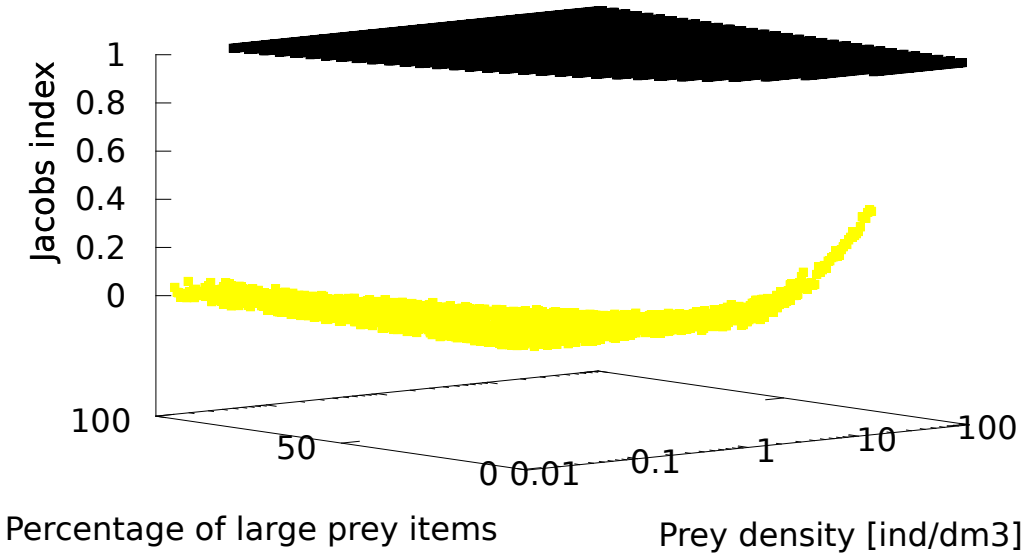
Active choice based on local in space information on prey distribution violates the classical Zero-One Rule. Prey items above the critical value are generally profitable, but are attacked only with some probability smaller than 1. This result can be clearly seen on Figure 3.12 - when the population consists of many small items and only few large ones, it is profitable to forage on both prey types. The frequency at which a small item is found attractive is, however, declining with increasing encounter rate.

3.2.9. Effect of predator's memory

3.2.9.1. Impact of short-term memory on foraging efficiency

The model can also be used to investigate the importance of predator's memory in the context of remembering location of prey items. We can address this problem by two quantitative methods. Firstly, assuming the predator has perfect memory, we may check in simulations how often it is profitable to turn back to capture a prey item currently outside of VFV. Secondly, by comparing possible net energy intake rates when foraging with perfect memory and with no memory about positions of encountered prey items, we

Figure 3.12: BOFM active selectivity index (3.11) of large prey category (2.5mm) in the population at density level N (x-axis) with p percent of small items - 1.5mm (y-axis). Yellow dots represent situations when capturing small prey items is profitable; black dots - situations when small prey are ignored.



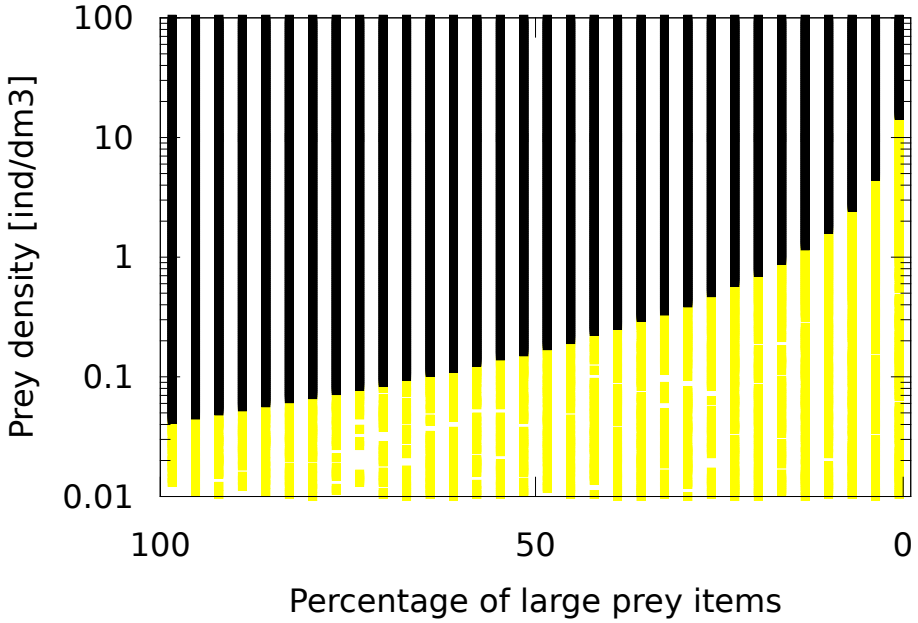
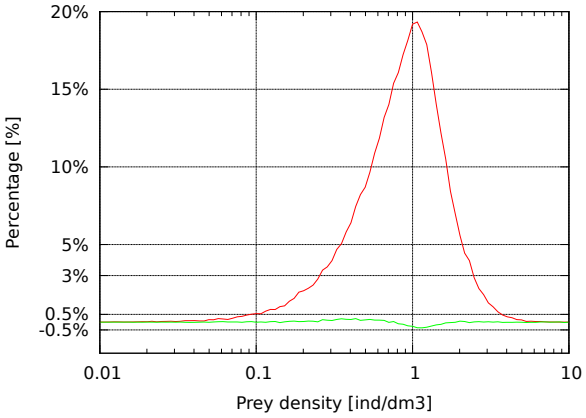


Figure 3.13: Significance of predator's memory: red line - percentage of captured prey items that wouldn't have been approached if predator didn't remember their locations; green line - improvement of net energy intake rate that was achieved thanks to perfect memory.



may assess by how much the onset of short-term memory increases predators evolutionary fitness, measured by the increase of the rate of net energy intake.

As can be seen in Figure 3.13 predators refer to their memory relatively often (up to 20% times) in the middle range of prey densities. It is intuitively clear that in low densities prey items outside of VFV are statistically too far to be profitable, while in high densities there is always a good choice of prey within VFV. However, it turns out that fitness improvement, understood as an average rate of net energy gain, resulting from the perfect memory of locations of all encountered prey items is negligible (at most 0.5%). We can therefore speculate, that there is no evolutionary pressure on aquatic predators to develop short-term memory even in the case where prey are immotile and hence short-term memory would precisely reflect their actual positions. In Section 3.2.9.2 we argue that long-term memory, concerning mean abundance of environment and structure of patches, can greatly affect foraging strategies as well as foraging efficiency in terms of energy intake.

3.2.9.2. Long-term memory and its impact on selectivity

A question of whether the predator is more size-selective in higher or in lower prey density appears in many different contexts. We believe that the answer greatly depends on the predator's long-term strategy, which can be either harvesting (optimizing efforts within a given prey abundance) or searching (moving through space in order to find a better habitat). Local information is insufficient for the decision-making process so we infer that the strategy is chosen based on long-term memory regarding heterogeneity of space. We have built two models to reflect both strategies: BOFM which is based solely on local information and predicts behavior in harvesting strategy while MOFM is a simplistic model of searching strategy.

In both models the selectivity index can be easily computed by inserting

$$\begin{aligned} p_i &= \frac{u(\Omega_i)}{u[0, s_{max}]}, \\ r_i &= \frac{\int_{\Omega_i} \alpha(s) r(s)^2 u(ds)}{\int_0^{s_{max}} \alpha(s) r(s)^2 u(ds)} \end{aligned}$$

into (3.10), where attack probability $\alpha(s)$ is either given by (3.15) in BOFM or (3.16) in MOFM. Notice that if $\alpha(s) = 0$ for the investigated size range Ω_i then $D_i = -1$ and also if Ω_i is the whole interval on which α equals 1, namely $\Omega_i = \{s : \alpha(s) = 1\}$, then $D_i = 1$. As an immediate consequence of these equations we infer that the selectivity is higher in the searching strategy than in the harvesting provided $v_a > v_0$. Indeed all the prey categories that are captured upon encounter in BOFM are also captured in MOFM as the condition

$$P(s, v_{opt}, 0) \geq \mathbb{E}P(\sigma, v_{opt}, \delta + r(\sigma))$$

implies

$$P(s, 0, 0) - \frac{A(v_a)}{T_h} \geq P_a.$$

The later follows from the fact that P_a being the anticipated rate of energy gain in a patch is bigger then that elsewhere thus

$$P_a \geq \mathbb{E}P(\sigma, v_{opt}, \delta + r(\sigma)).$$

In homogeneous environment (e.g. restricted in space) predators learn that harvesting is the optimal strategy. The comparison of selectivity in low and high densities within this strategy is presented in Figure 3.12.

In heterogeneous (e.g. patchy environment) predator forages using harvesting strategy in high density (within patches) and searching strategy in low density (elsewhere). In this case selectivity in low density does not depend on local abundance nor prey size-distribution. It results from the anticipated abundance of a patch reflected by the values of parameters v_a and P_a (compare Section 3.4). These parameters cannot be assessed based on local information and have to be sensed by the predator and kept in its memory. As fitness greatly depends on global foraging strategy, including searching for patches, it is allowed to infer that evolutionary changes favor development of long-term memory of patchy environment characteristics.

3.2.9.3. The shape of functional response

In Section 3.2.8.2 we obtained an approximation (3.14) of the capture rate in the case of low prey encounter rate. If v were a constant parameter and $\alpha(s)$ were a given function (independent of $u(\cdot)$) the functional response formula would exactly be the Holling type II function. However, in our optimal foraging model both α and v depend on prey size-distribution and overall food abundance, namely $\alpha(s) = \mathbb{1}_{[s_{min}, s_{max}]}(s)$ and $v = v_{cruis}$, where $\mathbb{1}_{[s_{min}, s_{max}]}(s)$ is a characteristic function equal to 1 if $s \in [s_{min}, s_{max}]$ and 0 elsewhere.

From Figure 3.12 we infer that, in the general case (for possibly high prey encounter rate), attack probability α ranges from 0 to 1 (assuming values not necessarily equal to 0 or 1) depending on both density and structure of prey population. High prey encounter rate also induces strong effect of variable distance to chosen prey items. This phenomenon is neglected in Holling model and thus the approximation of our model shows higher inaccuracy for high prey encounter rates (compare Figure 3.14). Asymptotic behavior of the functional response for high prey density is, however, easy to express in terms of formula (3.14). In the limit, all but the largest prey are ignored and the distance between consecutive chosen prey items is infinitesimal, allowing the predator to capture nearly one item per T_h time.

Functional responses computed with BOFM simulation and it's approximation by formula (3.14) for a prey population consisting of two size categories equal in number are depicted in Figure 3.14. The point of discontinuity corresponds to switching strategy between capturing both types of prey (lower prey densities) and capturing only the larger ones (higher prey densities). Despite that, the plot of the rate of net energy gain is continuous.

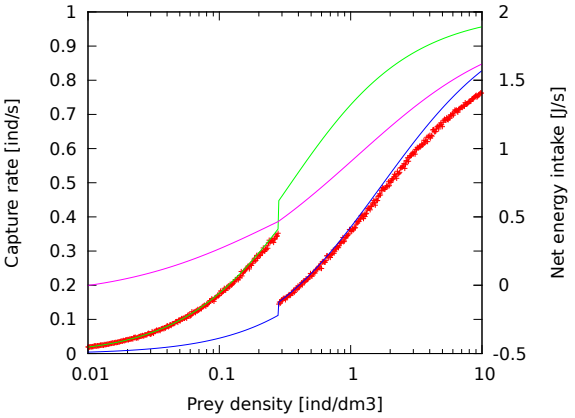
As noticed in Section 3.2.9.2, BOFM and, in consequence, the simulation results depicted in Figure 3.14 apply to situations in which the predator senses the homogeneity of the environment and optimizes its efforts within the habitat. Such situations include environments limited in space (small ponds), experimental systems, and patches. The model predicts existence of low-density refuge and, more precisely, a marginal prey density (0.01 $\frac{\text{ind.}}{\text{dm}^3}$) below which no items are captured.

The predictions are different in heterogeneous systems and homogeneous systems, where the space is so large that the predator is unaware of their homogeneity. In such cases, the predator's searching strategy in low density is be modeled by MOFM and harvesting strategy within patches by BOFM. The behavior in searching mode depends on the predator's long-term memory (reflecting 'knowledge' about the patchiness, density and structure of prey population), and thus experimental systems need to be carefully designed to ensure that the predator had enough time to train to forage in the tested environment. Predictions of functional response resulting from both strategies for a single predator BOFM and MOFM are shown in Figure 3.16. The predator decreases its capture rate in low densities (compared to optimal foraging - red line) in order to relocate to the patch faster. The velocity and expectations about the patch may vary from one individual predator to another, but each of them follows the searching strategy (MOFM) until the density meets its expectations, and switches to harvesting in higher abundance. We conclude that the marginal density, apparent in BOFM, does not exist in MOFM and, in consequence, in heterogeneous nor large-scale environments. The sigmoidal shape of functional response, however, results from the switch in strategies rather than unprofitability of foraging.

3.2.10. Implementation of the model

The model was implemented in C++11 language and all simulations were performed on x86_64 architecture, each running on a single core.

Figure 3.14: Functional response in BOFM in the case of two size categories equally distributed (red points) approximated by formula (3.14) with no selectivity (green line) and with selectivity for large prey (blue line). Dependence of net energy intake upon prey density (pink line) was shown on the right scale. Notice the log-scale on x-axis. The jumps in approximations (green and blue lines) result from discontinuity in velocity function.



2

Reactive distance is computed using the Newton–Raphson method. Computing cruising speed (Section 3.2.4) and optimal velocity (Section 3.2.5) is found by golden section search. In every time step of the simulation a finite section of the space around the predator is modeled (prey sizes and locations are stored in a data structure that contains information about a large ball around the predator). In order to eliminate boundary effects, every time when the virtual predator gets close to the border of its 'universe,' a set of new prey items is generated in the empty field that has not been visited before. The number of new prey items is drawn from Poisson distribution and their positions are drawn from uniform distribution (using random number generators from standard C++11 library).

3.3. Foraging in the framework of measure theory

In Section 3.2 foraging is characterized as a sequential process of capturing individual prey items. Such an approach allows incorporating most realistic assumptions and obtaining numerical results. For the purpose of farther modeling (e.g. model ling of prey population dynamics or its space distribution), however, it is more convenient to represent functional response as an operator on the space of size-distributions (i.e. $C : \mathfrak{M}^+[0, s_{max}] \to \mathfrak{M}^+[0, s_{max}]$), similarly as it was done in Section 3.2.8.2.

For a given process of capturing individual prey items it is natural to define a capture rate operator $C : \mathfrak{M}^+[0, s_{max}] \to \mathfrak{M}^+[0, s_{max}]$ as

$$C_{BOFM}[u](E) = \lim_{T\to\infty} \frac{\#\{\text{prey items of sizes restricted to } E \text{captured in BOFM in time } T\}}{T}.$$

General models such as individual-based BOFM or MOFM introduced in Section 3.2.5 are quite complex and difficult to analyze in the framework of operators on the space of measures. In particular, it is not clear whether the definition above is correct and the

Figure 3.15: Functional response in BOFM (red points) for uniformly distributed four size categories of prey (a) and 16 size categories (b) respectively. Dependence of net energy intake upon prey density (green) was shown on the right scale on both pictures. Dependence of s_{min} upon prey density in the case of 16 size categories (c). Notice the log-scale on x-axis.

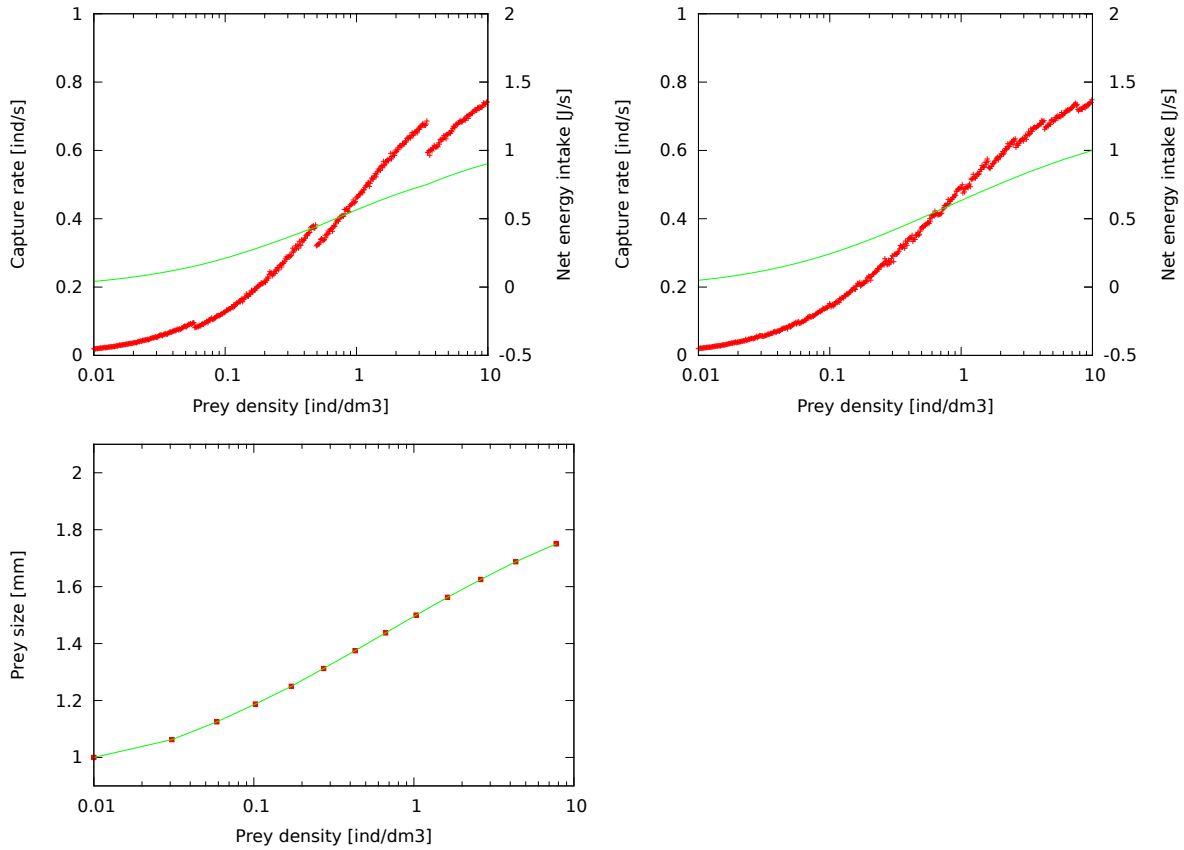
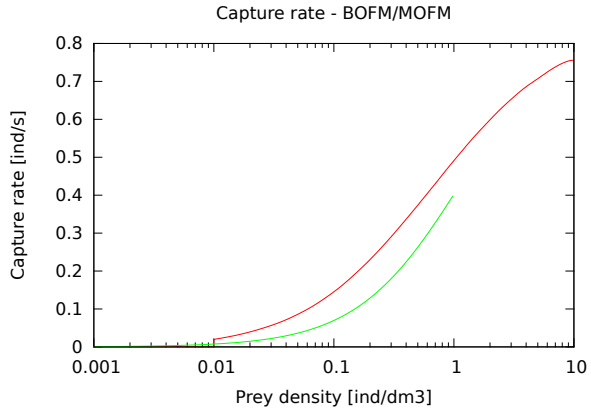


Figure 3.16: Functional responses in BOFM (red line) and MOFM (green line) for uniformly distributed continuous range of prey sizes between $1mm$ and $2mm$, and $v_a = 3\frac{dm}{s}$, and P_a reflecting expected rate of net energy gain in patch with abundance $1\frac{ind}{dm^3}$. The minimal size of captured prey items in MOFM is equal $1.73mm$.



conditions under which the limit exists are unknown. Basic properties of the processes such as decreased selectivity in higher densities or increased velocity in higher temperatures can be easily proved. On the contrary, natural questions arising from measure theory approach, such as Lipschitz continuity of C_{BOFM}, are almost impossible to solve. Nonetheless, in such cases a numerical study can be conducted and it proves to be useful for constructing simpler, yet accurate enough, models suitable for population dynamics.

The aim of this section is to define capture rate operator

$$C_{LOW} : \mathfrak{M}^+[0, s_{max}] \to \mathfrak{M}^+[0, s_{max}],$$

which is easier to analyze than C_{BOFM} and still capable of grasping all the important phenomena such as stable prey-predator interactions, variable predator velocity and passive selectivity. At this point we abandon the bottom-up approach of building more complex models on top of simpler ones following evident causal relations. Instead, we shall restrict our considerations to the domain of low population densities and following the lines of Section 3.2.8.2 we enhance the classical Holling formula to incorporate desired dependence on the whole size-structure of the population. The new model is justified by the comparison against BOFM.

Let

$$C_{LOW}\,[u] = \frac{\pi v[u] r^2 u}{1 + T_h \pi v[u] \int_0^{s_{max}} r^2(\sigma) u(d\sigma)},$$

for $v : \mathfrak{M}^+[0, s_{max}] \to \mathbb{R}$ being the maximizer of expected rate of net energy intake. Instead of employing the most general form of expected rate of net energy intake introduced in 3.2.4, namely $\mathbb{E}_{\sigma,\delta} P(\sigma, v, \delta + r(\sigma))$, we assume that $\delta >> r(\sigma)$ and $\frac{v}{\delta} >> T_h$ and for given u we derive

$$\mathbb{E}_{\sigma,\delta} P(\sigma, v, \delta) \approx \mathbb{E}_{\sigma} P(\sigma, v, \mathbb{E}\delta) \approx P(u, v) =$$

$$= \pi v \int_0^{s_{max}} r^2(\sigma) \left(e(\sigma) - A(v) - R(v) \frac{1}{v\pi \int_{s_0}^{s_{max}} r^2(\sigma) u(d\sigma)} \right) u(d\sigma) =$$

$$= \pi v \int_0^{s_{max}} r^2(\sigma) \left(e(\sigma) - A(v) \right) u(d\sigma) - R(v).$$

We also make a particular choice of functions A and R, namely let $A(v) = \frac{mv^2}{2}$ (compare Section 3.2.6) and let $R(v) = r_0 + r_1 v + r_2 v^2 + r_3 v^3$. Function $R(v)$ gives a good approximation of the respiration rate introduced in (3.4) for $v \in [0, 13]$ if v is measured in meters per second provided that $r_0 = 6.8 \cdot 10^{-3}$, $r_1 = 1.24 \cdot 10^{-3}$, $r_2 = 6.0 \cdot 10^{-5}$, $r_3 = 2.5 \cdot 10^{-5}$. The range of velocity is based on the experimental data.

3.4. Discussion

In this chapter a new, mechanistic, individual-based approach to modeling of visually foraging predators constantly searching for and capturing prey items in a prey population

with prescribed size structure. Prey items are assumed to be immobile and homogeneously distributed in 3-dimensional space. The model is based on the classical concept of optimal foraging but contrary to previous works, all aspects of predator's decisions are being subjected to optimization. Underlying assumptions on aquatic habitats and the limitations of predator's perception, described in Section 3.2.3, are somewhat idealized and may serve as a reference point for more particularized studies.

One of our main assumptions is inspired by results of experiments described in [29]. It concerns the ability of planktivorous fish to make decisions on capturing or ignoring individual prey basing on locally perceivable information as well as on globally assessed prey abundance. We claim that these two factors along with the prey's energy value, the predator's respiration rate and the amount of energy, $A(v)$, needed to accelerate after prey capture to velocity v determines the final choice of prey item. Empirical assessment of $A(v)$ is a challenging task indicating the direction of further studies.

Identifying the circumstances under which it is profitable for the predator to ignore a perceived individual prey is an important component of our model. Intuitively speaking, it precisely defines when a prey item is too small or the distance to a prey item is too large. The terms 'too small' and 'too large' always need a reference point and in our models it's either the average rate of net energy income characterizing given habitat (in harvesting strategy) or the anticipated energy (in searching strategy). The predator's ability to sense prey population density is assumed in BOFM (which applies to habitats with homogeneous in space prey distribution and patches). MOFM predicts the predator's behavior in an intermediate position between patches provided two parameters: searching velocity, v_a, and the anticipated net energy intake, P_a. In this paper we do not consider theoretical methods of evaluating the choice of parameters v_a and P_a. It is, however, intuitively clear that P_a should reflect the net energy intake achievable in the patch by harvesting strategy provided that there is no risk of starvation. We believe the optimal velocity v_a for a patchy environment can be computed using similar methods as in Section 3.2.4 when applied to the distribution of patches instead of prey items. In the long timescale, the predator's fitness is usually measured by the number of offspring or exhaustion time (the time until satiety falls to zero for the first time) rather than average net energy intake (eg. [6]). The difference between these models of fitness is particularly important when the danger of starvation is considerable and optimization of energy involves high risk. Such ideas give an alternative method of determining values v_a and P_a.

While in classical models of predation (such as Holling-type functional response) both predator's speed and selectivity are assumed, the approach used in this chapter allows for predicting these values. The comparison of results for BOFM and MOFM indicates that in the heterogeneous environment selectivity in high density stems from a different cause than selectivity does in low density. It is therefore important to distinguish 'relatively low density' and 'low density' when speaking of selectivity - the first term relates to heterogeneous environment while the latter to a homogeneous one.

The optimal foraging model developed in this article can be extended in many directions to take into account various processes related to foraging. Several factors have potentially high impact on foraging strategies: predator's degree of satiation, risk tolerance, dependence of risk upon light conditions, and sensitivity on light conditions coupled

with corresponding changes in prey recognition. The subject of this study is restricted to foraging of an individual predator, nonetheless conclusions can be used as a building block for further studies of phenomena such as population dynamics, vertical distribution of prey or patch exploitation. Our study is focused on predation itself and on its impact on the structure of prey population leaving aside other factors which affect its size and structure. Notice that the processes of predation and population growth are of different time scale and in the case of planktivorous fish, active foraging is restricted to a short time at dawn and dusk. The changes in population structure due to birth and natural death in this time may be neglected. Thus, per capita mortality predicted by our optimal foraging models can be used in more general structured population models to describe full population dynamics. We also believe that patch exploitation studies can be enriched by the observation arising from Figure 3.8 that higher abundance deters predators from patrolling larger areas.

In Section 3.2.1 the classical Holling disk equation is viewed from the perspective of optimal foraging theory. This approach allows us to predict the occurrence of low prey density refuge resulting from predator's negative rate of energy intake. An empirically testable conjecture, stating that in the presence of visually foraging predator a power law determines the relation between the density of a prey population and average prey size was formulated.

Two different types of predator's selectivity (passive, resulting from the immanent selectivity of predator's sight, and active, resulting from predator's choice) are often discussed in literature (see eg. [83]). The structure BOFM and MOFM enabled us to incorporate both ideas in one framework and therefore obtain realistic predictions for both low encounter rate (when passive selectivity plays a crucial role) and high encounter rate (when active selectivity becomes an important factor). Accuracy of predictions is additionally supported by the resemblance of functional response predicted by the model (Figure 3.16) and the experimental data (Figure 3.2).

Finally, a model based on Holling disk equation, enriched by reactive distance and energy balance models was introduced in Section 3.3 to reflect the functional response of a visual predator optimizing its cruising speed in low encounter rate. As observed in [29] aggregation of plankton in open ecosystems imposes higher risk of being captured on each prey item and in consequence is maladaptive. Also aggregational response of the predators is strong enough to eliminate patches on zooplankton. We infer that the model may be inaccurate in the general case of possibly high encounter rate, but for the aforementioned reasons is sufficient for modeling population dynamics in real habitats.

Final remarks

As a final note we present some unresolved issues and open problems related to this dissertation, that the author found particularly interesting.

Algorithms for computing flat distance

In Section 1.3.4 an algorithm for computing flat distance between two measures from $\mathfrak{M}^+_{d,N}(\mathbb{R})$ was presented. The complexity of this algorithm was proved to be $\mathcal{O}(N \log N)$. Can this result be improved?

1. Does there exist an algorithm for computing flat distance between two measures from $\mathfrak{M}^+_{d,N}(\mathbb{R})$ with linear complexity, $\mathcal{O}(N)$?

2. Does there exist a linear algorithm which, given two measures $\mu, \nu \in \mathfrak{M}^+_{d,N}(\mathbb{R})$, computes an upper bound for the flat metric, $\overline{\rho_F}(\mu, \nu)$, satisfying

$$\rho_F(\mu, \nu) \leq \overline{\rho_F}(\mu, \nu) \leq C \cdot \rho_F(\mu, \nu)$$

   for some constant C? What is the smallest constant C for which such algorithm exists?

Approximation theory for Radon measures

Theorem 70 provides an estimate of the flat distance between a Lipschitz continuous function, $f \in C^{0,1}[0,1]$, and its optimal N-step approximation, f^N. It turns out that $\rho_F(f, f^N) \leq C \cdot N^{-2}$ for some constant C. How does this result generalize to other classes of functions and their approximations?

1. Assume $f \in C[0,1]$ is only a continuous function and let f^N be its optimal N-step approximation. Does the following asymptotic behavior hold

$$\rho_F(f, f^N) = \mathcal{O}(N^{-2})?$$

2. Fix $f \in C^{0,1}[0,1]$, and let f^N be its optimal linear spline. Does $\rho_F(f, f^N) = \mathcal{O}(N^{-3})$, if N is the number of intervals on which f^N is linear?

Section 1.5.3 provides a method for approximating continuous functions, $f \in C[0,1]$, by discrete N-point measures. The length of the interval $[0,1]$ plays a crucial role in the reasoning, see Remark 38. How can this result be generalized to the case of arbitrary interval $[a,b]$ instead of $[0,1]$.

Transport equation

The tools used for the study of McKendrick-von Foerster equation can be generalized to transport equations. In this dissertation solutions are considered in the space of $(\mathfrak{M}^+(X), \rho_F)$, which is not a linear space. The linear space $(\mathfrak{M}(X), \rho_F)$, on the other hand is not complete.

1. How can elements of the Banach completion, $\overline{(\mathfrak{M}(X), \rho_F)}$, be characterized?
2. Can Theorem 79 be generalized to the space $\overline{(\mathfrak{M}(X), \rho_F)}$?
3. How can the methods of computing distances between measures be generalized to compute distances between elements of $\overline{(\mathfrak{M}(X), \rho_F)}$?

Model of zooplankton population Theorem 109 characterizes stationary state to the McKendrick-von Foerster equation with mortality resulting from the optimal foraging model.

1. Is the stationary state, characterized by Theorem 109, stable?
2. What is the rate of convergence to the stationary state?
3. What is the basin of attraction of the stationary state?

Optimal foraging model

In Chapter 3 a post-capture acceleration cost function, $A(v)$, was introduced to reflect the energy expense of predator when accelerating from a motionless state to velocity v. In this dissertation it was assumed that $A(v)$ is equal to predator's kinetic energy at velocity v, namely $\frac{mv^2}{2}$. How accurate is this estimation? An experimental study of predator's respiration rate during acceleration could answer this question and provide basis for more precise optimal foraging models.

Bibliography

[1] Abrams, P.A. 1982. Functional Responses of Optimal Foragers. American Naturalist, 120, 382-390.

[2] Ackleh A. S., Banks H. T., Deng K., 2001, A finite difference approximation for a coupled system of non-linear size-structured populations, Nonlinear Analysis 50 727-748.

[3] Aljetlawi, A.A., Sparrevik, E., Leonardsson, K., 2004. Prey-predator size-dependent functional response: derivation and rescaling to the real world. J. Animal Ecology. 73, 239-252.

[4] Ambrosio L., Crippa G., 2008, Existence, Uniqueness, Stability and Differentiability Properties of the Flow Associated to Weakly Differentiable Vector Fields in Transport Equations and Multi-D Hyperbolic Conservation Laws, Lect. Notes Unione Mat. Ital., Springer 5: 3-57.

[5] Ambrosio L., Gigli N., Savaré. G., 2005, Gradient Flows in Metric Spaces and in the Space of Probability Measures, Birkhäuser, ETH Lecture Notes in Mathematics.

[6] Barton, K., Hovestadt T., 2013. Prey density, value, and spatial distribution affect the efficiency of area-concentrated search. J. Theo. Biol. 316, 61-69.

[7] Bartosiewicz M., Jabłoński J., Kozłowski J., Maszczyk P., Brood space limitation of reproduction may explain growth after maturity in differently sized zooplankton, in preparation.

[8] Bolnick, D.I. and Ferry-Graham, L. A. 2002. Optimizing prey-capture behaviour to maximize expected net benefit. Evol. Ecology Research, 4, 843-855.

[9] Brannstrom A., Carlsson L., Simpson D., On the convergence of the escalator boxcar train, arXiv:1210.1444v1.

[10] Bressan A., 2000, Hyperbolic systems of conservation laws, Oxford Lecture Series in Mathematics and its Applications, vol. 20, Oxford University Press, Oxford, 2000.

[11] Carrillo J.A., Colombo R.M, Gwiazda P., Ulikowska A., 2012, Structured populations, cell growth and measure valued balance laws, J. Diff. Eq. 252: 3245–3277.

[12] Carrillo J.A, Francesco M. Di., Toscani G., 2007, Strict contractivity of the 2-Wasserstein distance for the porous medium equation by mass-centering, Proc. Amer. Math. Soc. 135: 353–363.

[13] Carrillo J. A., McCann R. J., Villani C., 2006, Contractions in the 2–Wasserstein length space and thermalization of granular media, Arch. Rational Mech. Anal., 179:217–264.

[14] Charnov, E.L., 1976. Optimal Foraging, the marginal Value Theorem.Theor. Popul. Biol.. 9, 141-151.

[15] De Roos A. D., 1989, Daphnids on a Train, Development and Application of A New Numerical Method for Physiologically Structured Population Models, Rijksuniversiteit te Leiden.

[16] De Roos A. D., 1997, A Gentle Introduction to Physiologically Structured Population Models, Structured-Population Models in Marine, Terrestrial, and Freshwater Systems Population and Community Biology Series Volume 18, pp 119-204.

[17] De Roos A. D., Persson L., 2002, Size-dependent life-history traits promote catastrophic collapses of top predators, PNAS, Vol. 99, No. 20.

[18] Dunbrack, R.L. and Giguere, L.A. 1987. Adaptive Responses to Accelerating Costs of Movement: A Bioenergetic Basis for the Type-III Functional Response. American Naturalist, 130, 147-160.

[19] Edmonds J., Karp R. M., 1972, Theoretical Improvements in Algorithmic Efficiency for Network Flow Problems, Journal of the ACM, Vol. 19, No. 2.

[20] Eggers, D. M., 1982. Planktivore preference by prey size. Ecology 63(2), 381-390.

[21] Emlen, J.M. 1966. The role of time and energy in food preference. American Naturalist 100: 611-617.

[22] Evans L.C., Gariepy R.F, 1992, Measure Theory and Fine Properties of Functions, CRC Press.

[23] Fortet R., Mourier B., 1953, Convergence de la repartition empirique vers la repartition theoretique, Ann. Sci. Ecole Norm. Sup. 70 , 267.

[24] Gangbo W., Mccann R. J., 1999, Shape recognition via Wasserstein distance, Applied Mathematics.

[25] Gause G.F., Smaragdova N.P., Witt A.A., 1936, Further studies of interaction between predators and prey, The Journal of Animal Ecology 5, 1–18.

[26] Gardner, M.B. 1981. Mechanisms of size selectivity by planktivorous fish: A test of hypotheses.. Ecology. 63(3), 571-578.

[27] Giske, J., Rosland, R., Berntsen, J., Fiksen Ø., 1997. Ideal free distribution of copepods under predation risk, Ecological modelling 95 45-59.

[28] Gliwicz, Z. M. 2001. Species-specific population-density thresholds in cladocerans? Hydrobiologia 442:291–300.

[29] Gliwicz Z. M., Maszczyk P., Jabłoński J., Wrzosek D., 2013, Patch exploitation by planktivorous fish and the concept of aggregation as an antipredation defense in zooplankton, Limnol. Oceanogr. 58: 1621–1639.

[30] Gliwicz Z. M., Szymańska E., Wrzosek D., 2010, Body size distribution in *Daphnia* populations as an effect of prey selectivity by planktivorous fish, Hydrobiologia, Volume 643, Issue 1, pp 5-19.

[31] Gondzio J., Terlaky T., 1996, A computational view of interior point methods, Advances in linear and integer programming, pp. 103–144.

[32] Gourcy M., 2007, A large deviation principle for 2D stochastic Navier–Stokes equation, Stochastic Processes and their Applications, Vol. 117, No. 7.

[33] Gliwicz Z. M., Wrzosek D., 2008, Predation-mediated coexistence of large- and small-bodied *Daphnia* at different food levels. American Naturalist 172, 358-374.

[34] Gwiazda P, Jablonski J, Marciniak-Czochra A, Ulikowska A, Analysis of particle methods for structured population models with nonlocal boundary term in the framework of bounded Lipschitz distance, arXiv:1309.2408.

[35] Gwiazda P., Lorenz T., Marciniak-Czochra A., 2010, A nonlinear structured population model: Lipschitz continuity of measure valued solutions with respect to model ingredients, J. Diff. Eq. 248: 2703–2735.

[36] Gwiazda P., Marciniak-Czochra A., 2010, Structured population models in metric spaces, J. Hyper. Diff. Eq. 7: 733–773.

[37] Hedenstrom, A., Alerstam, T.1995. Optimal Flight Speed of Birds. Phil. Trans. R. Soc. Lond. B 29. 348,471-487.

[38] Holling, C.S., 1959. Some characteristics of simple tpes of predation and parasitism, Canad. Entomol. 91.

[39] Jabłoński J., Approximation of Radon Measures in Flat Metric and Applications in Modelling, in preparation.

[40] Jablonski J., Marciniak-Czochra A., Efficient algorithms computing distances between Radon measures on R, arXiv:1304.3501.

[41] Jabłoński J., Wrzosek D., Functional response resulting from an optimal foraging model of a size-selective predator-harvester, submitted.

[42] Jacobs, J. 1974. Quantitative measurements of food selection; a modification of the forage ratio and Ivlev's selectivity index. Oecologia 14: 413–417.

[43] Jordan R., Kinderlehrer D., Otto F., 1998, The variational formulation of the Fokker-Planck equation, SIAM J. Math. Anal. 29: 1–17.

[44] Kinderlehrer D., Walkington N.J., 1999, Approximation of parabolic equations using the Wasserstein metric, Math. Mod. Num. Anal. 33: 837–852.

[45] Klein M., 1967, A Primal Method for Minimal Cost Flows with Applications to the Assignment and Transportation Problems, Management Science, Vol. 14, No. 3.

[46] Kostova T., 2003, An explicit third-order numerical method for size-structured population equations, Numer. Methods Partial Differential Equations 19(1): 1–21.

[47] Krivan, V., 2013, Behavioral refuges and predator-prey coexistence., J. Theor. Biol. in press.

[48] Krzyżanowski P., Wrzosek D., Wit D., 2006, Discontinuous Galerkin method for piecewise regular solutions to the nonlinear age-structured population model,Mathematical Biosciences, Vol. 203, No. 2.

[49] Lazarro, X., 1987. A review of planktivorous fishes: Their evolution, feeding behaviours, selctivities, and impacts. Hydrobiologia 146, 97-167.

[50] Lipman Y., Daubechies I., 2011, Conformal Wasserstein distances: Comparing surfaces in polynomial time, Advances in Mathematics, Vol. 227, No. 3.

[51] MacArtur, R.H. and Pianka., 1966. On optimal use of a patchy environment. American Naturalist 100: 603-609.

[52] Manatunge, J., Asaeda, T.,1990. Optimal foraging as the criteria of prey selection by two centrarchid fishes, Hydrobiologia 391, 223-240.

[53] Maszczyk, P., Gliwicz, M.Z., 2014. Selectivity by planktivorous fish at different prey densities, heterogeneities, and spatial scales, Limnol. Oceanogr. 59(1), 68-78.

[54] Maury B., Roudneff-Chupin A., Santambrogio F., 2010, A macroscopic crowd motion model of gradient flow type, Mathematical Models and Methods in Applied Sciences.

[55] Maury B., Roudneff-Chupin A., Santambrogio F., Venel J., 2011, Handling congestion in crowd motion modeling, Networks and Heterogeneous Media.

[56] McKendrick A. G., 1926, Application of mathematics to medical problems, Pro. Edinburgh Math. Soc., 44, pp. 98-130.

[57] Mittelbach, G.G. and Osenberg C.W.,1994. Using foraging theory to study trophic interactions. in D.J. Stouder, K.L.Fresh and R.J. Feller (eds.) Theory and Application in Fish Feeding Ecology. 45-59.

[58] Mueller-Merback H., 1966, An Improved Starting Algorithm for the Ford-Fulkerson Approach to the Transportation Problem, Management Science, Vol. 13, No. 1.

[59] Murdoch W. W., Oaten H., 1975, Predation and population stability. Adv. Ecol. Res. 9: 2- 125.

[60] Neunzert H., 1981, An introduction to the nonlinear Boltzmann-Vlasov equation, in Kinetic Theories and the Boltzmann Equation, Springer, Berlin, Lecture Notes in Math. 1048: 60–110.

[61] Oudre L., Jakubowicz, J., Bianchi P., Simon C., 2012, Classification of Periodic Activities Using the Wasserstein Distance, Biomedical Engineering, IEEE Transactions, Vol. 59, No. 6.

[62] Piccoli B., Rossi F., *On properties of the Generalized Wasserstein distance, arXiv:1304.7014.*

[63] Pflug G. Ch., Pichler A., 2011, Approximations for Probability Distributions and Stochastic Optimization Problems, International Series in Operations Research & Management Science Volume 163, pp 343-387.

[64] Pütter A., 1920, Studien über physiologische Ähnlichkeit VI. Wachstumsähnlichkeiten, Physiologie des Menschen und der Tiere, Vol. 180, No. 1.

[65] Pyke, G.H., Pulliam, H.R. Charnov, E.L. ,1977. Optimal foraging: a selective review of theory and tests. The Quarterly Rev. of Biology 52, 138-154.

[66] Pyke, G.H., 1981.Optimal travel speeds of animals. Am.Nat. 118, 475-487.

[67] Ranta E., Bengtsson J., McManus J., 1993, Growth, size and shape of Daphnia longispina, D. magna and D. pulex, Ann. Zool. Fennici 30:299-311.

[68] Rapoport E. O., Discerte Approximation of Continuous Measures and Some Applications, 2012, Journal of Applied and Industrial Mathematics, Vol. 6, No. 4, pp. 469-479.

[69] Solomon J., Rustamov R., Guibas L., Butscher A., 2014, Wasserstein Propagation for Semi-Supervised Learning, ICML 2014, Beijing.

[70] Stephens, D.W., Krebs, J.R., 1986. Foraging Theory, Princton Universiy Press, Princton, New Jersey.

[71] Sih, A. and Christensen, B. 2001. Optimal diet theory: when it works, and when and why does it fail?, Animal Behaviour 61, 379-390.

[72] Smith J. N. M., 1974, The food searching behavior of two European thrushes. I. Description and analysis of the search paths, Behavior 48, 276-302; II. The adaptiveness of the search patterns, Behavior 49, 1-61.

[73] Ulikowska A., 2012, An age-structured, two-sex model in the space of Radon measures: Well posedness, Kinetic and Related Models, 5: 873–900.

[74] Ulikowska A., 2013, PhD dissertation, http://www.mimuw.edu.pl/wiadomosci/aktualnosci/doktoraty/pliki/agnieszka_ulikowska/au-dok.pdf.

[75] Vallender S., Calculation of the Wasserstein Distance Between Probability Distributions on the Line, 1974, Theory of Probability & Its Applications, Vol. 18, No. 4 : pp. 784-786.

[76] Villani C., 2009, Optimal transport: old and new, Springer-Verlag, Berlin.

[77] Visser A.W., 2007.Motility of Zooplankton: fitness, foraging and predation. J. Plankton Res. 29, 447-461.

[78] Ware, D.M. 1975. Growth, methabolism and optimal swimming speed of a pelagic fish. J.Fish..Res. Board.Can. 32, 33-41.

[79] Ware, D.M.,1978. Bioenegetics of pelagic fish: theoretical change in swimming speed and ration with body size. J. Fish. Res. Board Can. 35, 220-228.

[80] Weaver N., 1999, Lipschitz Algebras , World Scientic Publishing Co. Pte. Ltd.

[81] Werner, E.E., Hall, D.J.,1974. Optimal foraging and the size selection of prey by the bluegill sunfish (Lepomis Macrochirus). Ecology 52, 1042-1052.

[82] Westdickenberg M., Wilkening J., 2010, Variational particle schemes for the porous medium equation and for the system of isentropic Euler equations Math. Mod. Num. Anal. 44: 133–166.

[83] Wetterer, J.K, Bishop, C. J., 1985. Planktivore prey selection: The reactive field volume model vs. the apparent size model. Ecology 66(2) 457-464.

[84] Yi-Te, L., Jiun-Hong C., Ling-Ling L., 2011, Prey selection of a shell-invading leech as predicted by optimal foraging theory with consumption success incorporated into estimation of prey profitability.