Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Learning a Sparse Representation of Barron Functions with the Inverse Scale Space Flow

Tjeerd Jan Heeringa * [1], Tim Roith [2], Christoph Brune [1], and Martin Burger [2,3]

[1]Mathematics of Imaging & AI, University of Twente, Enschede, The Netherlands.
[2]Helmholtz Imaging, Deutsches Elektronen-Synchrotron DESY, Hamburg 22607, Germany.
[3]Fachbereich Mathematik, Universität Hamburg, Hamburg 20146, Germany.

Abstract. This paper presents a method for finding a sparse representation of Barron functions. Specifically, given an L^2 function f, the inverse scale space flow is used to find a sparse measure μ minimising the L^2 loss between the Barron function associated to the measure μ and the function f. The convergence properties of this method are analysed in an ideal setting and in the cases of measurement noise and sampling bias. In an ideal setting the objective decreases strictly monotone in time to a minimizer with $\mathcal{O}(1/t)$, and in the case of measurement noise or sampling bias the optimum is achieved up to a multiplicative or additive constant. This convergence is preserved on discretization of the parameter space, and the minimizers on increasingly fine discretizations converge to the optimum on the full parameter space.

Keywords:
Barron Space,
Bregman Iterations,
Sparse Neural Networks,
Inverse Scale Space,
Optimization.

Article Info.:
Volume: 4
Number: 1
Pages: 48 - 88
Date: March/2025
doi.org/10.4208/jml.240123

Article History:
Received: 23/01/2024
Accepted: 25/12/2024

Communicated by:
Chenglong Bao

1 Introduction

Most neural networks contain a subnetwork with fewer parameters that performs equally well [36], and some of these subnetworks have been found to generalise equally or even better than their dense counterparts [28, 29]. However, it is a priori hard to determine which parameters of the network will be part of the subnetwork. Hence, various approaches have been developed for finding well performing sparse neural network. They fall roughly in three categories. The first is to add a term to the loss or regularizer that promotes sparsity. An example of this would be the least absolute shrinkage and selection operator (LASSO), in which a ℓ^1 regularizer is added [39]. The second is to train a network first and prune it afterwards, meaning weights are reduced with as little as possible influence on the performance [31]. The third is to start with a sparse architecture, and add or remove neurons during training [22].

One of the methods, which starts from a sparse architecture, is based on the Bregman iteration [33]. This method has been introduced and thoroughly analysed for imaging and compressed sensing [15, 17, 44]. The method works in these settings by progressively adding more detail to the reconstructed images and signals, respectively. A limitation of the original method is that it requires that often requires the problem to be convex.

*Corresponding author. t.j.heeringa@utwente.nl

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

However, adaptations of the method, e.g. the linearized variant in [5,13], where the loss is replaced by a first order approximation, allows for a successful application to neural networks. A major success of this method is that it is able to find an auto-encoder without ever explicitly defining an auto-encoder like architecture [12]. This shows that it has major potential for automatic neural network architecture design tasks.

1.1 Related work

Bregman iterations were introduced in [33] and further developed and analysed in [1,6,15, 17–19,43,44] as an algorithm to solve sparsity promoting regularisation tasks in computer vision. Linearized Bregman iterations as introduced in [18,44] can be seen as a generalization of the mirror descent algorithm [4,32] to the non-differentiable, convex case. More recently, variants of the original algorithm have been applied in the context of machine learning, see, e.g. [12,13,40,41].

Bregman iterations are the implicit Euler discretization of an inverse scale space flow. Going to the continuous limit has helped to find easy implementations for relatively complex functionals like the total variation functional, and has helped to obtain well-justified and simple stopping criteria [14]. In the finite-dimensional case of sparse regularization (and further generalizations) an exact time discretization can be found, which leads to efficient methods [15,30]. We refer to [6] for recent overview.

Similar to inverse scale space flow being the continuous limit of the Bregman iterations, we have that the Barron spaces are the continuous limit of shallow neural network. It was proven that Barron functions have bounded point evaluations [2,38], Barron functions can be approximated in L^p with rate $\mathcal{O}(m^{-1/p})$ [26], Barron spaces have a represented theorem [34] and that Barron spaces are a kind of integral reproducing kernel Banach spaces (RKBS), a Banach space analogue to reproducing kernel Hilbert spaces (RKHS) [2]. The spaces are parametrized by the activation function of the networks. The Barron spaces associated to most of the commonly used non-periodic activation are embedded in the Barron space with ReLU as activation function [27]. This Barron space together with the Barron spaces associated to the rectified power unit (RePU), the higher-order generalization of the ReLU, are strongly related to bounded variation (BV) spaces [26,34].

A fundamental open question in machine learning is how to find the best function representing your data. For Barron spaces, this means finding the best measure μ representing the Barron function f. Since the relation between μ and f is linear, this leads to a convex minimization problem. Based on an alternative representation of Barron functions in probability space, the authors in [42] formulated a Wasserstein gradient flow for this problem based on the ideas of [21]. Under several assumptions, including omnidirectional initial conditions and satisfying the Morse-Sard property, this leads to a unique solution π [42]. However, not all Barron functions satisfy the Morse-Sard property, placing a limit on the functions that can be represented with this approach [42]. Although this unique solution π represents the Barron function f, it is not necessarily the probability measure for f with the smallest semi-norm. In order to find sparse neural networks, there is a need for a method that minimizes this semi-norm as well.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

1.2 Background information

This section provides the relevant background information needed of Barron spaces and Bregman iterations.

1.2.1 Barron spaces

Fix $d \in \mathbb{N}$ and σ as an element of $\mathcal{C}^{0,1}(\mathbb{R})$ or the ReLU activation function $\max(0, x)$. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\Omega \subseteq \mathbb{R}^d \times \mathbb{R}$. Consider a probability measure $\rho \in \mathcal{P}(\mathcal{X})$, and define

$$K\mu(x) = \int_\Omega \sigma(a^\intercal x + b) d\mu(a, b) \tag{1.1}$$

for $\mu \in \mathcal{M}(\Omega)$. Barron space $\mathcal{B}_\sigma$ is the Banach space with functions of the form $f = K\mu$ for some $\mu \in \mathcal{M}(\Omega)$ and

$$\|f\|_{\mathcal{B}_\sigma} = \begin{cases} \inf\limits_{K\mu=f} \int_\Omega (1 + \|a\| + |b|) d|\mu|(a, b), & \sigma \in \mathcal{C}^{0,1}(\mathbb{R}), \\ \inf\limits_{K\mu=f} \int_\Omega (\|a\| + |b|) d|\mu|(a, b), & \sigma(x) = \mathrm{ReLU}(x). \end{cases} \tag{1.2}$$

The functions in Barron space can be seen as infinitely wide or continuous versions of shallow neural networks

$$f : \mathcal{X} \ \to \ \mathbb{R}, \quad x \ \mapsto \ \sum_{i=1}^{m} c_i \sigma\big(a_i^\intercal x + b_i\big) \tag{1.3}$$

with $c_i \in \mathbb{R}$ and $(a_i, b_i) \in \Omega$ [24]. Two embeddings are relevant for this work. They show that Barron functions are nice enough to enable proper convergence.

Proposition 1.1 (Barron is Lipschitz, [26, Theorem 3.3])**.** *If $\rho \in \mathcal{P}_1(\mathcal{X})$ is a probability measure with finite first moments, then we have $Lip(f) \leq Lip(\sigma)\|f\|_{\mathcal{B}_\sigma}$ for every $f \in \mathcal{B}_\sigma$.*

Proposition 1.2 (Barron L^p Embedding, [26, Theorem 3.7])**.** *If $\rho \in \mathcal{P}_q(\mathcal{X})$ is a probability measure with finite q-th moments, then $\mathcal{B}_\sigma \hookrightarrow L^p(\mathcal{X}, \rho)$ for all $1 \leq p \leq q$.*

1.2.2 Bregman iterations

Let $\mathcal{H}$ be some Banach space, $\mathcal{U}$ be a (closed subset of a) thereof, $f \in \mathcal{H}$, $J : \mathcal{U} \to \mathbb{R}$ be convex, lower semi-continuous and coercive, and $\mathcal{R}_f : \mathcal{U} \to \mathbb{R}$ be convex, bounded from below and Fréchet differentiable. The Bregman distance[1] between $u, v \in \mathcal{H}$ for $p \in \partial J(v)$ is given by

$$D_J^p(u, v) = J(u) - J(v) - \langle p \,|\, u - v \rangle. \tag{1.4}$$

[1]Although the Bregman distance is called a distance, it is in general neither symmetric nor does it satisfy the triangle inequality. It is also referred to as the Bregman divergence.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

The Bregman iterations

$$\begin{aligned} u_k &= \operatorname*{arg\,min}_{u\in\mathcal{U}} D_J^{p_{k-1}}(u,u_{k-1}) + \lambda\mathcal{R}_f(u), \quad u_0 = 0, \\ p_k &= p_{k-1} - \lambda\partial_u\mathcal{R}_f(u_k), \qquad\qquad\qquad p_0 = 0, \quad p_k \in \partial J(u_k) \end{aligned} \tag{1.5}$$

with design parameter $\lambda > 0$ are an iterative approximation algorithm for the bilevel minimization problem

$$\begin{aligned} &u^\dagger \in \operatorname*{arg\,min}_{u\in\mathcal{U}} J(u) \\ &\text{s.t.} \quad u \in \operatorname*{arg\,min}_{\bar{u}\in\mathcal{U}} \mathcal{R}_f(\bar{u}). \end{aligned} \tag{1.6}$$

The Bregman iterations converge monotonically to the optimal solution with worst case $\mathcal{O}(1/k)$ convergence [17].

The inverse scale space flow can be derived from (1.5) by taking the limit of $\lambda \searrow 0$. Before taking the limit, observe that (1.5) is equivalent to

$$u_k = \operatorname*{arg\,min}_{u\in\mathcal{U}\cap\partial J^*(p_k)} \frac{1}{\lambda}\big(J(u) - \langle p_{k-1} \,|\, u\rangle\big) + \mathcal{R}_f(u), \quad u_0 = 0 \tag{1.7a}$$

$$\frac{p_k - p_{k-1}}{\lambda} = -\partial_u\mathcal{R}_f(u_k), \qquad p_0 = 0. \tag{1.7b}$$

Note that usually (1.7b) has the subgradient constraint $p_k \in \partial J(\mu_k)$ instead of (1.7a) having $\partial J^*(p_k)$ as additional constraint. These two ways of writing the constraint are equivalent by Fenchel duality. In the limit of $\lambda \searrow 0$, (1.7b) can be seen as the Euler discretization of the flow equation

$$\partial_t p_t = -\partial_u\mathcal{R}_f(u_t), \quad p_0 = 0, \tag{1.8}$$

and (1.7a) will find a u_k minimizing $\mathcal{R}_f(u)$ whilst enforcing that $p_t \in \partial J(u_t)$ or equivalently $u_t \in \partial J^*(p_t)$ [11, 14, 37]. The inverse scale space is exactly this limit of $\lambda \searrow 0$ of the Bregman iterations, i.e. the dynamical process given by

$$u_t = \operatorname*{arg\,min}_{u\in\mathcal{U}\cap\partial J^*(p_t)} \mathcal{R}_f(u), \quad u_0 = 0, \tag{1.9a}$$

$$\partial_t p_t = -\partial_u\mathcal{R}_f(u_t), \qquad p_0 = 0. \tag{1.9b}$$

1.3 Our contribution

In this work, we study the convergence and error analysis of finding the smallest measure μ such that the Barron function $K\mu$ is close to f using the inverse scale space. This is the continuous and infinite dimensional version of finding a sparse shallow neural network approximating samples of f.

In particular, we consider the minimisation problem

$$\mu^{\text{opt}} = \operatorname*{arg\,min}_{\mu^+\in\mathcal{M}(\Omega)} J(\mu^+) \tag{1.10a}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

$$\text{s.t.} \quad \mu^\dagger \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}(\Omega)} \frac{1}{2}\|f - K\mu\|^2_{L^2(\rho)}, \tag{1.10b}$$

where J encodes the Barron norm and acts as regularizer and L_ρ is the adjoint of K. In Section 2 we define these operators more rigorously, and show that the associated inverse scale space is given by

$$\mu_t = \operatorname*{arg\,min}_{\mu \in \partial J^*(p_t)} \frac{1}{2}\|f - K\mu\|^2_{L^2(\rho)}, \quad u_0 = 0, \tag{1.11a}$$

$$\partial_t p_t = L_\rho(f - K\mu_t), \qquad p_0 = 0. \tag{1.11b}$$

The data function f and the data distribution ρ are instance dependent, and the convergence behaviour and the error analysis of (1.11) are dependent on these. In machine learning, measurements of f are noisy and the data sets always have a bias. Furthermore, computers are discrete beings. Hence, we analyse (1.11) in the following four cases:

1. Noiseless and unbiased case: We have access to f and sample from ρ.
2. Noisy case: We have access to f^δ with measurement noise instead to f, but we still want to find to a minimizer for f.
3. Biased case: We sample from ρ^ε with a sampling bias instead of from ρ, but we still want to find the minimizer for ρ.
4. Discretized case: The parameter space Ω is discretized and no longer continuous.

The first shows how well (1.11) can be when we manage to reduce noise and sampling bias to a minimum. The second shows how the methods deals with noise on the data function f. The third provides a novel perspective on learning methods. It shows how well the method deals with a bias in the sampling. In machine learning there is a large focus on computing the generalisation error of a method, i.e. how large is the error you make when you solve (1.10) with only n samples of ρ relative to using ρ in its entirety. This is one way of having a bias in the sampling. Another bias that one could have as the goal to classify animals based on images to determine whether they are suitable pets, but one has no images of fish. Our method captures both of these biases in one go. The last shows that the method behaves nicely when the parameter space Ω is discretized.

We show in Section 2 that the (1.11) is well-defined and determine its optimality conditions. After that we discuss the aforementioned four cases in Sections 3 to 6 respectively.

2 Inverse scale space flow for Barron spaces

In this section, we start by defining the necessary functionals and operators to write down the inverse scale space flow for Barron spaces. In Section 2.1, we show how to get from the general form of the inverse scale space in (1.9) to (2.3). Then, in Section 2.2, we show that this flow is well-defined. Last, in Section 2.4, we derive several optimality conditions for the flow that are needed for the proofs of the convergence rates later in this work.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Fix $d \in \mathbb{N}$. Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $\Omega \subseteq \mathbb{R}^{d+1}, \rho \in \mathcal{P}_2(\mathcal{X})$ be a probability measure with bounded second moment, $\sigma \in \mathcal{C}^{0,1}(\mathbb{R})$ or $\sigma(x) = \max(0, x)$, $V(a,b) = 1 + \|a\| + |b|$ and $f \in L^2(\rho)$, where we mean that $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ when we write $(a,b) \in \Omega$. Use these to define the operators

$$K : \mathcal{M}(\Omega) \to L^2(\mathcal{X}, \rho), \quad \mu \mapsto \left(x \mapsto \int_\Omega \sigma(a^\intercal x + b) d\mu(a,b) \right), \tag{2.1a}$$

$$L_\rho : L^2(\mathcal{X}, \rho) \to C(\Omega), \quad \phi \mapsto \left((a,b) \mapsto \int_{\mathcal{X}} \phi(x)\sigma(a^\intercal x + b) d\rho(x) \right), \tag{2.1b}$$

$$J : \mathcal{M}(\Omega) \to [0, \infty), \quad \mu \mapsto \int_\Omega V(a,b) d|\mu|(a,b), \tag{2.1c}$$

$$\mathcal{R}_f : \mathcal{M}(\Omega) \to [0, \infty), \quad \mu \mapsto \frac{1}{2}\|K\mu - f\|^2_{L^2(\mathcal{X},\rho)}. \tag{2.1d}$$

We consider the task of finding

$$\mu^{\text{opt}} \in \operatorname*{arg\,min}_{\mu^\dagger \in \mathcal{M}(\Omega)} J(\mu^\dagger) \tag{2.2a}$$

$$\text{s.t.} \quad \mu^\dagger \in \operatorname*{arg\,min}_{\mu \in \mathcal{M}(\Omega)} \mathcal{R}_f(\mu). \tag{2.2b}$$

The constraint in (2.2b) says that we are looking for a measure μ such that $K\mu$ represents the $L^2(\rho)$ projection of f onto the Barron space, and (2.2a) highlights that we want the measure that induces the Barron norm. We will search for the measure μ^{opt} using the inverse scale space flow. The flow corresponding to (2.2) is given by

$$\mu_t = \operatorname*{arg\,min}_{\mu \in \partial J^*(p_t)} \mathcal{R}_f(\mu), \quad u_0 = 0, \tag{2.3a}$$

$$\partial_t p_t = L_\rho(f - K\mu_t), \quad p_0 = 0. \tag{2.3b}$$

In the following, we will assume that every $\mu^\dagger$ we refer to has finite $J(\mu^\dagger)$.

2.1 Derivation of the inverse scale space flow for Barron spaces

To derive the inverse scale space flow for Barron spaces, we start with (1.5) and (1.9). These imply that the Bregman iterations and associated inverse scale space flow for (2.2) are given by the iterative process

$$\mu_k = \operatorname*{arg\,min}_{\mu \in \mathcal{M}(\Omega)} D_J^{p_{k-1}}(\mu, \mu_{k-1}) + \lambda \mathcal{R}_f(\mu), \quad \mu_0 = 0, \tag{2.4a}$$

$$p_k = p_{k-1} - \lambda \partial_\mu \mathcal{R}_f(\mu_k), \qquad p_0 = 0, \quad p_k = \partial J(\mu_k), \tag{2.4b}$$

and the dynamical system

$$\mu_t = \operatorname*{arg\,min}_{\mu \in \mathcal{M}(\Omega) \cap \partial J^*(p_t)} \mathcal{R}_f(\mu), \quad \mu_0 = 0, \tag{2.5a}$$

$$\partial_t p_t = -\partial_\mu \mathcal{R}_f(\mu_t), \qquad p_0 = 0, \tag{2.5b}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

respectively. First, observe that $\partial J^*(p_t) \subseteq \mathcal{M}(\Omega)$. This shows that (2.5a) and (2.3a) match. Before we show that (2.5b) is the same as (2.3b), we show that L_ρ is in fact the adjoint of K.

Lemma 2.1. *The adjoint L_ρ is given by K, i.e. $L_\rho^\star = K$.*

Proof. Let $\phi \in L^2(\mathcal{X}, \rho)$ and $\mu \in \mathcal{M}(\Omega)$, then, by Fubini-Tonelli

$$\begin{aligned}\langle K\mu \,|\, \phi \rangle_{L^2(\rho)} &= \int_\Omega \int_{\mathcal{X}} \sigma(a^\intercal x + b) d\rho(x) \phi(x) d\mu(a,b) \\ &= \int_\Omega \int_{\mathcal{X}} \phi(x) \sigma(a^\intercal x + b) d\rho(x) d\mu(a,b) \\ &= \langle \mu \,|\, L_\rho \phi \rangle_{\mathcal{M}(\Omega)}.\end{aligned}$$

From the definition of the adjoint it follows that $L_\rho^\star = K$. □

Note that K is the adjoint for all L_ρ with $\rho \in \mathcal{P}_2(\mathcal{X})$, but that the difference between the various L_ρ is the inner product used.

Proposition 2.1. *The variational derivative of $\mathcal{R}_f$ is given by*

$$\partial_\mu \mathcal{R}_f(\mu) = L_\rho(K\mu - f). \tag{2.6}$$

Proof. Observe that

$$\begin{aligned}&\lim_{\|\nu\|_{\mathcal{M}(\Omega)} \to 0} \frac{|\mathcal{R}_f(\mu+\nu) - \mathcal{R}_f(\mu) - \langle \partial_\mu \mathcal{R}_f(\mu) \,|\, \nu \rangle_{\mathcal{M}(\Omega)}|}{\|\nu\|_{\mathcal{M}(\Omega)}} \\ = &\lim_{\|\nu\|_{\mathcal{M}(\Omega)} \to 0} \frac{\left| \|K(\mu+\nu) - f\|^2_{L^2(\rho)}/2 - \|K\mu - f\|^2_{L^2(\rho)}/2 - \langle K^*(K\mu - f) \,|\, \nu \rangle_{\mathcal{M}(\Omega)} \right|}{\|\nu\|_{\mathcal{M}(\Omega)}} \\ \leq &\lim_{\|\nu\|_{\mathcal{M}(\Omega)} \to 0} \frac{\left| \|K\nu\|^2_{L^2(\rho)}/2 + \langle K\mu - f \,|\, K\nu \rangle_{L^2(\rho)} - \langle K^*(K\mu - f) \,|\, \nu \rangle_{\mathcal{M}(\Omega)} \right|}{\|\nu\|_{\mathcal{M}(\Omega)}} \quad \text{(triangle ineq.)} \\ = &\lim_{\|\nu\|_{\mathcal{M}(\Omega)} \to 0} \frac{\left| \|K\nu\|^2_{L^2(\rho)}/2 \right|}{\|\nu\|_{\mathcal{M}(\Omega)}} \quad \text{(def. of adjoint)} \\ \leq &\lim_{\|\nu\|_{\mathcal{M}(\Omega)} \to 0} \frac{1}{2} \|K\|^2_{op} \|\nu\|_{\mathcal{M}(\Omega)} = 0.\end{aligned}$$

Hence,

$$\partial_\mu \mathcal{R}_f(\mu) = K^*(K\mu - f). \tag{2.7}$$

Combining Lemma 2.1 with (2.7) finishes the proof. □

This shows that (2.5b) is indeed the same as (2.3b), and thus that (2.5) is the same as (2.3).

2.2 Existence

To show that the inverse scale space flow of (2.3) has a solution, we use a theorem by Brezis [9, Theorem 3.1]. This theorem establishes that the differential inclusion equation

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

$$\partial_t r_t + Br_t \in 0 \tag{2.8}$$

given some initial condition $r_0 \in \mathrm{dom}(B) := \{r \in H \mid Br \neq \emptyset\}$ has a solution. Here, B is a maximally monotone, possibly nonlinear and possibly multivalued operator over a Hilbert space H. We show that for a suitably chosen maximal operator B, the solution to (2.8) exists, that for suitably chosen operator $\tilde{\mathcal{B}}$ the inverse scale space flow of (2.3) can be written in the form (2.8) using $\tilde{\mathcal{B}}$, and that the solution to the former flow satisfies the dynamics of the latter flow. Thereby establishing the existence of a solution to (2.3).

These suitably chosen operators $\tilde{B}$ and B are

$$A : \mathcal{C}(\Omega) \to \mathcal{M}(\Omega), \quad p \mapsto \operatorname*{arg\,min}_{\mu \in \partial \chi_{\{\|\cdot\|_\infty \leq 1\}}(p)} \mathcal{R}_f(\mu), \tag{2.9a}$$

$$\tilde{B} : L^2(\rho) \to L^2(\rho), \quad r \mapsto KA\left(V^{-1} L_\rho r\right) - f, \tag{2.9b}$$

$$B : L^2(\rho) \to L^2(\rho), \quad r \mapsto K\partial J^*(L_\rho r) - f. \tag{2.9c}$$

First, we show (2.8) with operator B has a solution. For this, we use that it is maximal monotone.

Lemma 2.2. *The operator B is maximal monotone.*

Proof. J^* is the Fenchel dual of J. Hence, J^* is lower semi-continuous, convex and proper. L_ρ is a bounded linear operator, so $J^* \circ L_\rho$ is also lower semi-continuous, convex and proper. Thus, $r \mapsto \partial J^*(L_\rho r)$ is maximal monotone [10]. Subtracting a constant from a maximal monotone operator preserves maximal monotonicity, so B is maximal monotone. The proof is complete. □

This means the operator B satisfies the requirements for Brezis, and we thus have a solution.

Proposition 2.2. *For every $r_0 \in \mathrm{dom}(B)$ there exists a unique function $r : [0, \infty) \to L^2(\rho)$ such that*

1. *r satisfies* (2.8) *for almost every $t \in (0, \infty)$ with initial condition r_0,*
2. *$r_t \in \mathrm{dom}(B)$ for all $t > 0$,*
3. *r_t is Lipschitz continuous on $[0, \infty)$ with $\|\partial_t r_t\|_{L^\infty([0,\infty);L^2(\rho))} \leq \|B^\circ(r_0)\|_{L^2(\rho)}$,*
4. *r is right differentiable for all $t \in (0, \infty)$ and $\partial_t^+ r_t + B^\circ(r_t) = 0$ for all $t \in (0, \infty)$,*
5. *$t \mapsto B^\circ(r_t)$ is right continuous and $t \mapsto \|B^\circ(r_t)\|$ non-increasing,*

where

$$B^\circ(r_t) = \operatorname*{arg\,min}_{r \in B(r_t)} \|r\|_{L^2(\rho)}. \tag{2.10}$$

Proof. See [9, Theorem 3.1]. □

This does not show that (2.3) has a solution yet, since this satisfies (2.8) with the operator $\tilde{B}$ whereas (2.3) satisfies (2.8) with the operator B. The former about $\tilde{B}$ follows from the following lemma.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Lemma 2.3. *Eq.* (2.3) *can be written as*

$$\partial_t r_t + \tilde{B}(r_t) = 0, \quad r = 0. \tag{2.11}$$

Proof. Substituting (2.9a) into (2.3) gives

$$\partial_t p_t = L_\rho\big(f - KA\big(V^{-1}p_t\big)\big), \quad p_0 = 0. \tag{2.12}$$

Replacing p_t with $L_\rho r_t$ gives us

$$L_\rho \partial_t r_t = L_\rho\big(f - KA\big(V^{-1}L_\rho r_t\big)\big), \quad r_0 = 0. \tag{2.13}$$

Since L_ρ is a bounded linear operator and thus continuous, r must satisfy

$$\partial_t r_t = f - KA\big(V^{-1}L_\rho r_t\big), \quad r_0 = 0, \tag{2.14}$$

or equivalently

$$\partial_t r_t + KA\big(V^{-1}L_\rho r_t\big) - f = 0, \quad r_0 = 0. \tag{2.15}$$

Substituting (2.9b) into (2.15) gives (2.11). □

We use the listed properties of the solution from Proposition 2.2 to show that the solution r to (2.8) with the operator B agrees with (2.8) with the operator $\tilde{B}$. This implies that there is a solution to (2.3).

Proposition 2.3. *Eq.* (2.3) *has a solution for every* μ_0 *and* p_0 *satisfying* $\mu_0 = A(V^{-1}L_\rho r_0)$ *and* $p_0 = L_\rho r_0$ *for some* $r_0 \in \mathrm{dom}(B)$. *In particular,* (2.3) *has a solution for* $\mu_0 = 0$ *and* $p_0 = 0$.

Proof. Let r be the solution from Proposition 2.2 with initial condition $r_0 \in \mathrm{dom}(B)$. Since

$$J^* = \chi_{\{\|V^{-1}\cdot\|_\infty \le 1\}}, \tag{2.16}$$

we have that

$$\begin{aligned} B^\circ(r_t) &= \operatorname*{arg\,min}_{x \in B(r_t)} \|x\|_{L^2(\rho)} \\ &= K\Big(\operatorname*{arg\,min}_{\mu \in \partial J^*(L_\rho r_t)} \|K\mu - f\|_{L^2(\rho)}\Big) - f \\ &= KA\big(V^{-1}L_\rho r_t\big) - f = \tilde{B}(r_t). \end{aligned} \tag{2.17}$$

So in fact, r also solves (2.8) with $\tilde{B}$, which has the same solution as (2.3) by Lemma 2.3. What remains is to map the solution r to μ and p using

$$\mu_t := A\big(V^{-1}L_\rho r_t\big), \quad p_t := L_\rho r_t.$$

The proof is complete. □

Remark 2.1. Note that this μ_t is not unique in general. Since the difference between non-uniqueness is from the null space of K, this does not impact any of the later statements.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

2.3 Regularity

The regularity that Proposition 2.2 puts on the solution r carries over to μ and p.

Proposition 2.4. *$\mu \in L^\infty([0,\infty), \mathcal{M}(\Omega))$ and $p \in \mathcal{W}^{1,\infty}([0,\infty), \mathcal{C}(\Omega))$.*

Proof. Recall from point 3 of Proposition 2.2 that

$$\|\partial_t r\|_{L^\infty([0,\infty),L^2(\rho))} \leq \left\|B^\circ\big(r(0)\big)\right\|_{L^2(\rho)} \leq \|f\|_{L^2(\rho)}. \tag{2.18}$$

This implies that

$$\|K\mu_t - f\|_{L^2(\rho)} = \|\partial_t r\|_{L^2(\rho)} \leq \|f\|_{L^2(\rho)}, \tag{2.19}$$

$$\|r_t\|_{L^2(\rho)} \leq \int_0^t \|\partial_s r_s\|_{L^2(\rho)} ds \leq t\|f\|_{L^2(\rho)}. \tag{2.20}$$

We will use this in the norm bounds for both μ and p.

For the regularity of p, observe that

$$\|L_\rho\|_{L^2(\rho)\to\mathcal{C}(\Omega)} = \|K\|_{\mathcal{M}(\Omega)\to L^2(\rho)} < \infty \tag{2.21}$$

by Lemma 2.1 and Proposition 1.2. Since $\partial_t p_t = L_\rho \partial_t r_t$, $p_t = L_\rho r_t$ and $r_t \in L^2(\rho)$, we have

$$\begin{aligned} \|p_t\|_{\mathcal{C}(\Omega)} &= \|L_\rho r_t\|_{\mathcal{C}(\Omega)} \leq \|L_\rho\|_{L^2(\rho)\to\mathcal{C}(\Omega)}\|r_t\|_{L^2(\rho)} \\ &\leq t\|L_\rho\|_{L^2(\rho)\to\mathcal{C}(\Omega)}\|f\|_{L^2(\rho)}, \end{aligned} \tag{2.22}$$

$$\begin{aligned} \|\partial_t p_t\|_{\mathcal{C}(\Omega)} &= \|L_\rho \partial_t r_t\|_{\mathcal{C}(\Omega)} \leq \|L_\rho\|_{L^2(\rho)\to\mathcal{C}(\Omega)}\|\partial_t r_t\|_{L^2(\rho)} \\ &\leq \|L_\rho\|_{L^2(\rho)\to\mathcal{C}(\Omega)}\|f\|_{L^2(\rho)} \end{aligned} \tag{2.23}$$

by (2.20), (3) of Proposition 2.3 and (2.21). Hence, $p \in \mathcal{W}^{\infty,1}([0,T), \mathcal{C}(\Omega))$ with

$$\|p\|_{\mathcal{W}^{1,\infty}([0,T),\mathcal{C}(\Omega))} \leq \max(1,t)\|L_\rho\|_{L^2(\rho)\to\mathcal{C}(\Omega)}\|f\|_{L^2(\rho)}. \tag{2.24}$$

For the regularity of μ, observe that

$$\begin{aligned} \|\mu_t\|_{\mathcal{M}(\Omega)} &\leq J(\mu_t) \\ &= \langle p_t \,|\, \mu_t\rangle_{\mathcal{M}(\Omega)} && \text{(Fenchel duality)} \\ &= \langle r_t \,|\, K\mu_t\rangle_{L^2(\rho)} \\ &\leq \|r_t\|_{L^2(\rho)}\|K\mu_t\|_{L^2(\rho)} && \text{(Cauchy-Schwarz)} \\ &= \|r_t\|_{L^2(\rho)}\|K\mu_t - f + f\|_{L^2(\rho)} \\ &\leq \|r_t\|_{L^2(\rho)}\big(\|K\mu_t - f\|_{L^2(\rho)} + \|f\|_{L^2(\rho)}\big) && \text{(Triangle ineq.)} \\ &\leq 2\|r_t\|_{L^2(\rho)}\|f\|_{L^2(\rho)} && \text{(Eq. (2.19))} \\ &\leq 2t\|f\|^2_{L^2(\rho)}. && \text{(Eq. (2.20))} \end{aligned}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Hence, $\mu \in L^\infty([0,T), \mathcal{M}(\Omega))$ with

$$\|\mu\|_{L^\infty([0,T),\mathcal{M}(\Omega))} \leq 2T\|f\|^2_{L^2(\rho)}. \tag{2.25}$$

Since the solution r is unique and the shown regularity holds for all $T > 0$, we can extend the regularity to the interval $[0, \infty)$. □

2.4 Optimality conditions

We have now proven the existence and regularity of the solutions to (2.3). In this section, we will have a look at some of the conditions that must hold for the optimal solution. In particular, the orthogonality condition, the first-order optimality condition and the source condition.

We first consider the orthogonality condition. This is a necessary condition, not a sufficient condition. This condition is equivalent to the first-order optimality condition for (2.2b).

Proposition 2.5 (Orthogonality Condition)**.**

$$L_\rho(f - K\mu^\dagger) = 0. \tag{2.26}$$

Proof. For $\mu^\dagger$ to be a minimizer of $\mathcal{R}_f$, it must hold that

$$\partial_\mu \mathcal{R}_f(\mu^\dagger) = 0. \tag{2.27}$$

Recall from Lemma 2.1 that

$$\partial_\mu \mathcal{R}_f(\mu) = L_\rho(f - K\mu). \tag{2.28}$$

Substituting (2.28) into (2.27) finishes the proof. □

The second condition we consider is the first-order optimality condition. This is again a necessary condition and not a sufficient condition.

Proposition 2.6 (First-Order Optimality Condition)**.**

$$\langle \partial_t p_t \,|\, \mu_t - \gamma \rangle \geq 0 \tag{2.29}$$

holds for all $t > 0$ and $\gamma \in \partial J^*(p_t)$.

Proof. For μ_t to minimize the right-hand side of (2.3a) it must satisfy the first-order optimality conditions for the right-hand side. Hence, μ_t must satisfy

$$\langle \partial \mathcal{R}_f(\mu_t) \,|\, \gamma - \mu_t \rangle \geq 0 \tag{2.30}$$

for all $\gamma \in \partial J^*(p_t)$. Substituting (2.5b) into (2.30) gives (2.29). □

J^* is a characteristic function, so ∂J^* is the normal cone given by

$$\begin{aligned} \partial J^*(p_t) &= \partial \chi_{\{\|V^{-1}\cdot\|_\infty \leq 1\}}(p_t) \\ &= \left\{ \mu \in \mathcal{M}(\Omega) \mid \langle \mu \,|\, q - p_t \rangle \leq 0,\ \forall\, q : \|V^{-1} q\|_\infty \leq 1 \right\}. \end{aligned} \tag{2.31}$$

In particular, $0 \in \partial J^*(p_t)$ for all $t \geq 0$. This gives the following corollary.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Corollary 2.1.

$$\langle \partial_t p_t \,|\, \mu_t \rangle \geq 0 \tag{2.32}$$

holds for all $t \geq 0$.

The third condition we consider is the source condition. It is satisfied by $\mu^\dagger$ if there exists a $\phi \in L^2(\mathcal{X}, \rho)$ such that

$$K^\star \phi \in \partial \int_\Omega V(a,b) d|\cdot|(\mu^\dagger). \tag{2.33}$$

Here, ϕ is called the Lagrangian multiplier, since it is the multiplier for the Lagrangian

$$\text{Lagrangian}(\mu, \phi) = J(\mu) - \langle \phi \,|\, K\mu - f \rangle_{L^2(\rho)}. \tag{2.34}$$

Hence, satisfying the source condition is akin to the existence of a Lagrange multiplier [16]. The following proposition provides an alternative representation for (2.33).

Proposition 2.7 (Source Condition)**.** *The source condition is satisfied by* $\mu^\dagger$ *if there exists a* $\phi \in L^2(\mathcal{X}, \rho)$ *such that*

$$L\phi(a,b) = V(a,b)\,\text{sgn}\{\mu^\dagger\} \quad \mu^\dagger \quad a.e., \tag{2.35}$$

and

$$|L\phi(a,b)| \leq V(a,b) \tag{2.36}$$

for all $(a,b) \in \Omega$.

Proof. We repeat the steps of Bredies in [8, around Eq. (4.1)], which in turn in based on [16, below Definition 1]. From the definition of the subdifferential it follows that (2.33) can only be satisfied when

$$\langle K^\star \phi \,|\, \nu \rangle_{\mathcal{M}(\Omega)} - \int_\Omega V(a,b) d|\nu| \leq \langle K^\star \phi \,|\, \mu^\dagger \rangle_{\mathcal{M}(\Omega)} - \int_\Omega V(a,b) d|\mu^\dagger| \tag{2.37}$$

for all $\nu \in \mathcal{M}(\Omega)$. Since

$$\langle K^\star \phi \,|\, \nu \rangle_{\mathcal{M}(\Omega)} = \langle \phi \,|\, K\nu \rangle_{L^2(\rho)} = \langle L_\rho \phi \,|\, \nu \rangle_{\mathcal{M}(\Omega)} \tag{2.38}$$

by the definition of the adjoint and Lemma 2.1, (2.37) is equivalent to

$$\langle L_\rho \phi \,|\, \nu \rangle_{\mathcal{M}(\Omega)} - \int_\Omega V(a,b) d|\nu| \leq \langle L_\rho \phi \,|\, \mu^\dagger \rangle_{\mathcal{M}(\Omega)} - \int_\Omega V(a,b) d|\mu^\dagger|. \tag{2.39}$$

Eq. (2.39) must also hold when we take the supremum of the left-hand side

$$\sup_{\nu \in \mathcal{M}(\Omega)} \langle L_\rho \phi \,|\, \nu \rangle_{\mathcal{M}(\Omega)} - \int_\Omega V(a,b) d|\nu| \leq \langle L_\rho \phi \,|\, \mu^\dagger \rangle_{\mathcal{M}(\Omega)} - \int_\Omega V(a,b) d|\mu^\dagger|. \tag{2.40}$$

Every measure $\nu \in \mathcal{M}(\Omega)$ has a polar decomposition such that

$$d\nu(a,b) = \text{sgn}\{\nu\}(a,b) d|\nu|(a,b). \tag{2.41}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

This allows us to write (2.40) as

$$\sup_{\nu \in \mathcal{M}(\Omega)} \langle L_\rho \phi - \operatorname{sgn}\{\nu\} V \,|\, \nu \rangle_{\mathcal{M}(\Omega)} \leq \langle L_\rho \phi \operatorname{sgn}\{\mu^\dagger\} - V \,|\, |\mu^\dagger| \rangle_{\mathcal{M}(\Omega)}. \tag{2.42}$$

The right-hand side is bounded, so must the left-hand side. If $L_\rho \phi(a,b) > V(a,b)$ for some $(a,b) \in \Omega$, then the left-hand side can be made arbitrarily large by concentrating a large positive ν around that value. Similarly, if $L_\rho \phi(a,b) < -V(a,b)$ for some $(a,b) \in \Omega$, then the left-hand side can be made arbitrarily large by concentrating a large negative ν around that value. Hence, $L_\rho \phi$ must satisfy

$$|L_\rho \phi(a,b)| \leq V(a,b). \tag{2.43}$$

Inserting this bound into (2.42) gives

$$0 = \sup_{\nu \in \mathcal{M}(\Omega)} \langle L_\rho \phi - \operatorname{sgn}\{\nu\} V \,|\, \nu \rangle_{\mathcal{M}(\Omega)} \leq \langle L_\rho \phi \operatorname{sgn}\{\mu^\dagger\} - V \,|\, |\mu^\dagger| \rangle_{\mathcal{M}(\Omega)} \leq 0. \tag{2.44}$$

Hence,

$$L_\rho \phi = V \operatorname{sgn}\{\mu^\dagger\}, \quad \mu^\dagger \quad \text{a.e.}. \tag{2.45}$$

The proof is complete. □

Note that the source condition described in Proposition 2.7 implies that μ_t must vanish on the set

$$\Omega_t^0 = \{(a,b) \in \Omega \mid -V(a,b) < p_t(a,b) < V(a,b)\}. \tag{2.46}$$

3 Idealized setting

In this section, we prove that both the L^2 loss $\mathcal{R}_f(\mu_t)$ and the Bregman distance $D_J^{p_t}(\mu^\dagger, \mu_t)$ decrease monotonically to the optimum value in an ideal setting. The rate at which both of them decrease is of order $\mathcal{O}(1/t)$. This rate is independent of the input dimension d.

Theorem 3.1 (Ideal Case). *$\mathcal{R}_f(\mu_t)$ is decreasing in time with bound*

$$\mathcal{R}_f(\mu_t) \leq \mathcal{R}_f(\mu^\dagger) + \frac{J(\mu^\dagger)}{t}, \quad t > 0 \quad \textit{a.e.}, \tag{3.1}$$

$$\partial_t D_J^{p_t}(\mu^\dagger, \mu_t) \leq 0, \quad t \geq 0 \quad \textit{a.e.} \tag{3.2}$$

with equality only when μ_t minimizes $\mathcal{R}_f$. Moreover, if $\phi \in L^2(\mathcal{X}, \rho)$ is the function such that the source condition of $\mu^\dagger$ is satisfied, then

$$D_J^{p_t}(\mu^\dagger, \mu_t) \leq \frac{\|\phi\|_{L^2(\rho)}^2}{2t} \tag{3.3}$$

for almost every $t \geq 0$.

First, we will show the rate of change of the L^2 loss $\mathcal{R}_f(\mu_t)$ and the Bregman distance $D_J^{p_t}(\mu^\dagger, \mu_t)$ under ideal conditions.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Lemma 3.1. *$\mathcal{R}_f(\mu_t)$ is decreasing in time.*

Proof. This follows directly from Proposition 2.2 point 5. □

Lemma 3.2.

$$\partial_t D_J^{p_t}(\mu^\dagger, \mu_t) \leq \langle \partial_t p_t \,|\, \mu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} \leq \mathcal{R}_f(\mu^\dagger) - \mathcal{R}_f(\mu_t) \leq 0 \tag{3.4}$$

holds for almost every $t \geq 0$.

Proof. Recall from the Fenchel duality that

$$J(\mu_t) = J(\mu_t) + J^*(p_t) = \langle p_t \,|\, \mu_t \rangle_{\mathcal{M}(\Omega)}. \tag{3.5}$$

Hence,

$$\begin{aligned}
\partial_t D_J^{p_t}(\mu^\dagger, \mu_t) &= \partial_t \big(J(\mu^\dagger) - J(\mu_t) - \langle p_t \,|\, \mu^\dagger - \mu_t \rangle_{\mathcal{M}(\Omega)} \big) \\
&= \partial_t \big(J(\mu^\dagger) - \langle p_t \,|\, \mu^\dagger \rangle_{\mathcal{M}(\Omega)} \big) && \text{(Eq. (3.5))} \\
&= \langle \partial_t p_t \,|\, -\mu^\dagger \rangle_{\mathcal{M}(\Omega)} \\
&\leq \langle \partial_t p_t \,|\, \mu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Corollary 2.1)} \\
&\leq \mathcal{R}_f(\mu^\dagger) - \mathcal{R}_f(\mu_t) && (-\partial_t p_t \in \partial \mathcal{R}_f(\mu_t)) \\
&\leq 0. && (\mu^\dagger \text{ minimizer})
\end{aligned}$$

The proof is complete. □

Proposition 3.1. *For all $t \geq 0$, it holds that*

$$\partial_t D_J^{p_t}(\mu^\dagger, \mu_t) < 0, \tag{3.6}$$

when

$$\|f - K\mu_t\|_{L^2(\rho)} > \|f - K\mu^\dagger\|_{L^2(\rho)} \tag{3.7}$$

as well as when

$$\|K\mu^\dagger - K\mu_t\|_{L^2(\rho)} > 0. \tag{3.8}$$

Proof. Eq. (3.7) holds if and only if

$$\mathcal{R}_f(\mu^\dagger) < \mathcal{R}_f(\mu_t). \tag{3.9}$$

The combination of (3.9) and (3.4) proves the first statement. For the second statement, observe that

$$\begin{aligned}
\partial_t D_J^{p_t}(\mu^\dagger, \mu_t) &\leq \langle \partial_t p_t \,|\, \mu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Lemma 3.2)} \\
&= \langle L_\rho(f - K\mu_t) \,|\, \mu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Eq. (2.3))} \\
&= \langle L_\rho(f - K\mu_t) - L_\rho(f - K\mu^\dagger) \,|\, \mu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Proposition 2.5)} \\
&= \langle L_\rho(K\mu^\dagger - K\mu_t) \,|\, \mu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)}
\end{aligned}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

$$
\begin{aligned}
&= \langle K\mu^\dagger - K\mu_t \,|\, K\mu_t - K\mu^\dagger \rangle_{L^2(\rho)} && \text{(Lemma 2.1)} \\
&= -\|K\mu^\dagger - K\mu_t\|^2_{L^2(\rho)}.
\end{aligned}
$$

Clearly, this is strictly negative when (3.8) is satisfied. □

Lemmas 3.2 and 3.1 show that under ideal conditions the Bregman distance and the population loss respectively are decreasing, and Proposition 3.1 shows that this decrease is strict. We will now use these to show that the Bregman distance and the population loss converge and give a rate at which they do that.

Proposition 3.2. *If $\mu^\dagger$ satisfies the source condition through $\phi \in L^2(\rho)$, then*

$$
D_J^{p_t}(\mu^\dagger, \mu_t) \leq \frac{\|\phi\|^2_{L^2(\rho)}}{2t} \tag{3.10}
$$

for almost every $t > 0$.

Proof. Define

$$
\partial_t e_t = K\mu^\dagger - K\mu_t, \quad e_0 = 0, \tag{3.11}
$$

and

$$
p^\dagger = L_\rho \phi. \tag{3.12}
$$

Observe that

$$
\partial_t p_t = L_\rho \partial_t e_t, \quad p_0 = 0 = L_\rho e_0. \tag{3.13}
$$

With this we obtain

$$
\begin{aligned}
&\partial_t \left(\frac{1}{2} \|e_t - \phi\|^2_{L^2(\rho)} \right) \\
&= \langle \partial_t e_t \,|\, e_t - \phi \rangle_{L^2(\rho)} \\
&= \langle K\mu^\dagger - K\mu_t \,|\, e_t - \phi \rangle_{L^2(\rho)} && \text{(Eq. (3.11))} \\
&= \langle L_\rho(e_t - \phi) \,|\, \mu^\dagger - \mu_t \rangle_{\mathcal{M}(\Omega)} && \text{(Lemma 2.1)} \\
&= \langle p_t - p^\dagger \,|\, \mu^\dagger - \mu_t \rangle_{\mathcal{M}(\Omega)} && \text{(Eqs. (3.13), (3.12))} \\
&= -\left(D^{p_t}(\mu^\dagger, \mu_t) + D^{p^\dagger}(\mu_t, \mu^\dagger) \right).
\end{aligned}
$$

Hence,

$$
\partial_t \left(\frac{1}{2} \|e_t - \phi\|^2_{L^2(\rho)} \right) + D^{p_t}(\mu^\dagger, \mu_t) \leq 0.
$$

Integrating from 0 to t gives

$$
\int_0^t D^{p_s}(\mu^\dagger, \mu_s) ds + \frac{1}{2} \|e_t - \phi\|^2_{L^2(\rho)} - \frac{1}{2} \|e_0 - \phi\|^2_{L^2(\rho)} \leq 0. \tag{3.14}
$$

Therefore,

$$
D^{p_t}(\mu^\dagger, \mu_t) = \frac{1}{t} \int_0^t D^{p_t}(\mu^\dagger, \mu_t) ds
$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

$$
\begin{aligned}
&= \frac{1}{t}\int_0^t D^{p_s}(\mu^\dagger, \mu_s)ds + \frac{1}{t}\int_0^t \int_s^t \partial_\tau D^{p_\tau}(\mu^\dagger, \mu_\tau)d\tau ds && \text{(Fund. th. of calc.)}\\
&\leq \frac{1}{t}\int_0^t D^{p_s}(\mu^\dagger, \mu_s)ds && \text{(Lemma 3.2)}\\
&\leq -\frac{1}{2t}\|e_t - \phi\|^2_{L^2(\rho)} + \frac{1}{2t}\|e_0 - \phi\|^2_{L^2(\rho)} && \text{(Eq. (3.14))}\\
&\leq \frac{1}{2t}\|e_0 - \phi\|^2_{L^2(\rho)}\\
&= \frac{1}{2t}\|\phi\|^2_{L^2(\rho)}. && \text{(Eq. (3.13))}
\end{aligned}
$$

The proof is complete. □

Proposition 3.3. *We have*

$$\mathcal{R}_f(\mu_t) \leq \mathcal{R}_f(\mu^\dagger) + \frac{J(\mu^\dagger)}{t} \tag{3.15}$$

for almost every $t > 0$.

Proof. Observe that

$$
\begin{aligned}
&D_J^{p_t}(\mu^\dagger, \mu_t) - (t-s)\big(\mathcal{R}_f(\mu^\dagger) - \mathcal{R}_f(\mu_t)\big)\\
&= D_J^{p_t}(\mu^\dagger, \mu_t) - \int_s^t \big(\mathcal{R}_f(\mu^\dagger) - \mathcal{R}_f(\mu_t)\big)d\tau\\
&\leq D_J^{p_t}(\mu^\dagger, \mu_t) - \int_s^t \big(\mathcal{R}_f(\mu^\dagger) - \mathcal{R}_f(\mu_\tau)\big)d\tau && \text{(Lemma 3.1)}\\
&\leq D_J^{p_t}(\mu^\dagger, \mu_t) - \int_s^t \partial_\tau D_J^{p_\tau}(\mu^\dagger, \mu_\tau)d\tau && \text{(Lemma 3.2)}\\
&= D_J^{p_s}(\mu^\dagger, \mu_s). && \text{(Fund. th. of calc.)}
\end{aligned}
$$

Hence, we obtain after rewriting

$$
\begin{aligned}
\mathcal{R}_f(\mu_t) &\leq \mathcal{R}_f(\mu^\dagger) + \frac{D_J^{p_s}(\mu^\dagger, \mu_s) - D_J^{p_t}(\mu^\dagger, \mu_t)}{t-s}\\
&\leq \mathcal{R}_f(\mu^\dagger) + \frac{D_J^{p_s}(\mu^\dagger, \mu_s)}{t-s} && (D_J^{p_t}(\mu^\dagger, \mu_t) \geq 0)\\
&\leq \mathcal{R}_f(\mu^\dagger) + \frac{D_J^{p_s}(\mu^\dagger, \mu_s)}{t} && (0 \leq s < t)\\
&\leq \mathcal{R}_f(\mu^\dagger) + \frac{D_J^{p_0}(\mu^\dagger, \mu_0)}{t} && \text{(Lemma 3.2)}\\
&= \mathcal{R}_f(\mu^\dagger) + \frac{J(\mu^\dagger)}{t}.
\end{aligned}
$$

The proof is complete. □

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

4 Measurement noise

In this section, we prove that with noise on the measurements, the method will converge with $\mathcal{O}(1/t)$ to the solution that best fits the noisy data. If the noise is small enough, then it will at first get closer to the noiseless data, too. After some time, the method will start to get close to the solution for the noisy data and will start moving away from the solution for the noiseless data. The point at which this transition is of the order of the noise, and suggest that the method should be stopped early in the presence of measurement noise.

In the remainder of the work, we consider f^δ to be some perturbation of f such that

$$\|f^\delta - f\|^2_{L^2(\rho)} \leq \delta \tag{4.1}$$

with $\delta > 0$. When using f^δ instead of f, the flow in (2.3) changes. For this section, we will keep referring to the solution based on f with μ and p whilst we will refer to the solution based on f^δ with ν and q.

Theorem 4.1 (Measurement Noise)**.** *We have*

$$\partial_t D^{p_t}(\mu^\dagger, \nu_t) \leq \frac{\delta^2}{4}, \quad t \geq 0 \quad \text{a.e.}, \tag{4.2}$$

$$\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) < 0, \quad t \geq 0 \quad \text{a.e.}, \tag{4.3}$$

when

$$\|f - K\nu_t\|_{L^2(\rho)} > \delta + \|f - K\mu^\dagger\|_{L^2(\rho)} \tag{4.4}$$

as well as when

$$\|K\mu^\dagger - K\nu_t\|_{L^2(\rho)} > \delta. \tag{4.5}$$

Moreover, if $\mu^\dagger$ satisfies the source condition through $\phi \in L^2(\mathcal{X}, \rho)$, then

$$D_J^{q_t}(\mu^\dagger, \nu_t) \leq \frac{1}{2t}(\|\phi\|_{L^2(\rho)} + \delta t)^2 + \frac{\delta^2 t}{8} \tag{4.6}$$

for almost every $t > 0$.

To prove this, observe that the flow for f^δ has the same properties as the flow for f.

Lemma 4.1. *$\mathcal{R}_{f^\delta}(\nu_t)$ is decreasing in t.*

Proof. Swapping the role of f and f^δ, i.e. considering f to be a perturbation of f^δ, implies that $\mathcal{R}_{f^\delta}(\nu_t)$ should behave the same as $\mathcal{R}_f(\mu_t)$ from Lemma 3.1. Thus, $\mathcal{R}_{f^\delta}(\nu_t)$ is decreasing in t. □

Lemma 4.1 shows that the inverse scale space converges with f^δ, but it does not tell us how close it will get to the best solution for f.

Lemma 4.2.

$$\partial_t D^{p_t}(\mu^\dagger, \nu_t) \leq \frac{\delta^2}{4} \tag{4.7}$$

holds for all $t \geq 0$.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Proof. This follows from

$$\begin{aligned}
&\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) \\
&\leq \langle \partial_t q_t \,|\, \nu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Lemma 3.2)} \\
&= \big\langle L\big(f^\delta - K\nu_t\big) \,\big|\, \nu_t - \mu^\dagger \big\rangle_{\mathcal{M}(\Omega)} && \text{(Eq. (2.3a))} \\
&= \big\langle L\big(f^\delta - K\nu_t\big) - L_\rho(f - K\mu^\dagger) \,\big|\, \nu_t - \mu^\dagger \big\rangle_{\mathcal{M}(\Omega)} && \text{(Proposition 2.5)} \\
&= \big\langle f^\delta - f + K\mu^\dagger - K\nu_t \,\big|\, K\nu_t - K\mu^\dagger \big\rangle_{L^2(\rho)} && \text{(Lemma 2.1)} \\
&= \big\langle f^\delta - f \,\big|\, K\nu_t - K\mu^\dagger \big\rangle_{L^2(\rho)} - \langle K\nu_t - K\mu^\dagger \,|\, K\nu_t - K\mu^\dagger \rangle_{L^2(\rho)} \\
&\leq \|f^\delta - f\|_{L^2(\rho)} \|K\nu_t - K\mu^\dagger\|_{L^2(\rho)} - \|K\nu_t - K\mu^\dagger\|_{L^2(\rho)}^2 && \text{(Cauchy-Schwarz)} \\
&\leq \frac{1}{4}\|f^\delta - f\|_{L^2(\rho)}^2 && \text{(Young's product ineq.)} \\
&\leq \frac{\delta^2}{4}.
\end{aligned}$$

The proof is complete. □

Proposition 4.1. *We have*

$$\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) < 0 \tag{4.8}$$

for all $t \geq 0$, *when*

$$\big\|f^\delta - K\nu_t\big\|_{L^2(\rho)} > \delta + \|f - K\mu^\dagger\|_{L^2(\rho)} \tag{4.9}$$

as well as when

$$\|K\mu^\dagger - K\nu_t\|_{L^2(\rho)} > \delta. \tag{4.10}$$

Proof. For the first statement observe that

$$\begin{aligned}
&\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) \\
&\leq \langle \partial_t q_t \,|\, \nu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Lemma 3.2)} \\
&= \big\langle L\big(f^\delta - K\nu_t\big) \,\big|\, \nu_t - \mu^\dagger \big\rangle_{\mathcal{M}(\Omega)} && \text{(Eq. (2.3a))} \\
&= \big\langle f^\delta - K\nu_t \,\big|\, K\nu_t - K\mu^\dagger \big\rangle_{L^2(\rho)} && \text{(Lemma 2.1)} \\
&= \big\langle f^\delta - K\nu_t \,\big|\, K\nu_t - f^\delta + f^\delta - f + f - K\mu^\dagger \big\rangle_{L^2(\rho)} \\
&= -\big\|f^\delta - K\nu_t\big\|_{L^2(\rho)}^2 + \big\langle f^\delta - K\nu_t \,\big|\, f^\delta - f + f - K\mu^\dagger \big\rangle_{L^2(\rho)} \\
&\leq -\big\|f^\delta - K\nu_t\big\|_{L^2(\rho)}^2 + \|f^\delta - f + f - K\mu^\dagger\|_{L^2(\rho)} \|K\nu_t - K\mu^\dagger\|_{L^2(\rho)} && \text{(Cauchy-Schwarz)} \\
&\leq -\big\|f^\delta - K\nu_t\big\|_{L^2(\rho)}^2 + \big(\delta + \|f - K\mu^\dagger\|_{L^2(\rho)}\big)\big\|f^\delta - K\nu_t\big\|_{L^2(\rho)}. && \text{(Triangle ineq., Eq. (4.1))}
\end{aligned}$$

Clearly, this is strictly negative when (4.9) is satisfied.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

For the second statement recall from the proof of Lemma 4.2 that

$$\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) \leq \|f^\delta - f\|_{L^2(\rho)} \|K\nu_t - K\mu^\dagger\|_{L^2(\rho)} - \|K\nu_t - K\mu^\dagger\|_{L^2(\rho)}^2.$$

Clearly, this is strictly negative when (4.10) is satisfied. □

From Proposition 4.1 and Lemma 4.2 it follows that the Bregman distance $D_J^{q_t}(\mu^\dagger, \nu_t)$ is guaranteed to converge until $\mathcal{R}_{f^\delta}(\nu_t)$ is close to $\mathcal{R}_f(\mu^\dagger)$. We know from Lemma 4.1 that $\mathcal{R}_{f^\delta}(\nu_t)$ will go to a minimum of $\mathcal{R}_{f^\delta}$. So we expect the Bregman distance $D_J^{q_t}(\mu^\dagger, \nu_t)$, unlike the Bregman distance $D_J^{q_t}(\mu^\dagger, \mu_t)$, to not vanish. The following proposition exemplifies this.

Proposition 4.2. *If $\mu^\dagger$ satisfies the source condition through $\phi \in L^2(\rho)$, then*

$$D_J^{p_t}(\mu^\dagger, \nu_t) \leq \frac{1}{2t}\big(\|\phi\|_{L^2(\rho)} + \delta t\big)^2 + \frac{\delta^2 t}{8} \tag{4.11}$$

for almost every $t \geq 0$.

Proof. Define

$$\partial_t e_t = f^\delta - K\nu_t + K\mu^\dagger - f, \quad e_0 = 0. \tag{4.12}$$

Observe that

$$\partial_t q_t = L_\rho \partial_t e_t, \quad q_0 = 0 = L_\rho e_0. \tag{4.13}$$

Using this definition of e_t we obtain

$$\begin{aligned}
&\partial_t \left(\frac{1}{2} \|e_t - \phi\|_{L^2(\rho)}^2 \right) \\
&= \langle \partial_t e_t \,|\, e_t - \phi \rangle_{L^2(\rho)} \\
&= \left\langle f^\delta - K\nu_t + K\mu^\dagger - f \,\middle|\, e_t - \phi \right\rangle_{L^2(\rho)} && \text{(Eq. (4.12))} \\
&= \left\langle f^\delta - f \,\middle|\, e_t - \phi \right\rangle_{L^2(\rho)} + \langle K\mu^\dagger - K\nu_t \,|\, e_t - \phi \rangle_{L^2(\rho)} \\
&\leq \|f^\delta - f\|_{L^2(\rho)} \|e_t - \phi\|_{L^2(\rho)} + \langle K\mu^\dagger - K\nu_t \,|\, e_t - \phi \rangle_{L^2(\rho)} && \text{(Cauchy-Schwarz)} \\
&\leq \delta \|e_t - \phi\|_{L^2(\rho)} + \langle K\mu^\dagger - K\nu_t \,|\, e_t - \phi \rangle_{L^2(\rho)} && \text{(Eq. (4.1))} \\
&= \delta \|e_t - \phi\|_{L^2(\rho)} + \langle L_\rho(e_t - \phi) \,|\, \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} && \text{(Lemma 2.1)} \\
&= \delta \|e_t - \phi\|_{L^2(\rho)} + \langle q_t - p^\dagger \,|\, \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} && \text{(Eq. (4.13), } p^\dagger := L_\rho(\phi)) \\
&= \delta \|e_t - \phi\|_{L^2(\rho)} - \langle q_t - p^\dagger \,|\, \nu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)}.
\end{aligned}$$

Since

$$0 \leq D_J^{p_t}(\mu^\dagger, \nu_t) + D_J^{p^\dagger}(\nu_t, \mu^\dagger) = \langle q_t - p^\dagger \,|\, \nu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)}, \tag{4.14}$$

where the inequality stems from that q_t and $p^\dagger$ are from the subgradients $\partial J(\nu_t)$ and $\partial J(\mu^\dagger)$ respectively, we obtain

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

$$\partial_t \left(\frac{1}{2} \|e_t - \phi\|^2_{L^2(\rho)} \right) \leq \delta \|e_t - \phi\|_{L^2(\rho)}. \tag{4.15}$$

Solving this for $\|e_t - \phi\|_{L^2(\rho)}$ gives

$$\|e_t - \phi\|_{L^2(\rho)} \leq \|e_0 - \phi\|_{L^2(\rho)} + \delta t = \|\phi\|_{L^2(\rho)} + \delta t. \tag{4.16}$$

Hence,

$$\begin{aligned} &\partial_t \left(\frac{1}{2} \|e_t - \phi\|^2_{L^2(\rho)} \right) + D_J^{p_t}(\mu^\dagger, \nu_t) \\ &\leq \delta \|e_t - \phi\|_{L^2(\rho)} - D_J^{p^\dagger}(\nu_t, \mu^\dagger) \\ &\leq \delta \|\phi\|_{L^2(\rho)} + \delta^2 t. \end{aligned} \tag{4.17}$$

By integrating both sides of the equation, we obtain

$$\begin{aligned} &\int_0^t D_J^{p_s}(\mu^\dagger, \nu_s) ds + \frac{1}{2} \|e_t - \phi\|^2_{L^2(\rho)} \\ &\leq \frac{1}{2} \|\phi\|^2_{L^2(\rho)} + \delta \|\phi\|_{L^2(\rho)} t + \frac{1}{2} \delta^2 t^2 \\ &= \frac{1}{2} \left(\|\phi\|^2_{L^2(\rho)} + \delta t \right)^2. \end{aligned} \tag{4.18}$$

Therefore,

$$\begin{aligned} D_J^{p_t}(\mu^\dagger, \mu_t) &= \frac{1}{t} \int_0^t D_J^{p_t}(\mu^\dagger, \mu_t) ds \\ &= \frac{1}{t} \int_0^t D_J^{s_t}(\mu^\dagger, \mu_s) + \int_s^t \partial_\tau D_J^{p_\tau}(\mu^\dagger, \mu_\tau) d\tau ds && \text{(Fund. th. of calc.)} \\ &\leq \frac{1}{t} \int_0^t D_J^{s_t}(\mu^\dagger, \mu_s) + \frac{\delta^2}{4} \int_s^t d\tau ds && \text{(Lemma 4.2)} \\ &= \frac{1}{t} \int_0^t D_J^{s_t}(\mu^\dagger, \nu_s) + \frac{\delta^2}{4} (t - s) ds \\ &= \frac{1}{t} \int_0^t D_J^{s_t}(\mu^\dagger, \nu_s) ds + \frac{\delta^2}{8} t \\ &\leq \frac{1}{t} \left(\frac{1}{2} \left(\|\phi\|^2_{L^2(\rho)} + \delta t \right)^2 - \|e_t - \phi\|^2_{L^2(\rho)} \right) + \frac{\delta^2}{8} t && \text{(Eq. (4.18))} \\ &\leq \frac{1}{2t} \left(\|\phi\|_{L^2(\rho)} + \delta t \right)^2 + \frac{\delta^2}{8} t. \end{aligned}$$

The proof is complete. □

Proposition 4.2 shows us that we should not continue to $t \to \infty$, but should stop earlier. In particular, the bound for (4.11) is lowest for $t(\delta) = \mathcal{O}(\delta^{-1})$.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

5 Biased sampling

In this section, we prove that a bias in the sampling gives a similar behaviour as noisy measurements. However, the terms and bounds differ depending on how the biased sampling is expressed. We consider sampling expressed in terms of a condition on either the Radon-Nikodym derivative or the Wasserstein-1 distance.

For the remainder of this work, we consider $\rho^\varepsilon \in \mathcal{P}_2(\mathcal{X})$ to be some perturbation of the true distribution $\rho \in \mathcal{P}_2(\mathcal{X})$, also with bounded second moment. We assume that $f \in L^2(\rho) \cap L^2(\rho^\varepsilon)$. For this section, we will keep referring to the solution based on ρ with μ and p whilst we will refer to the solution based on ρ^ϵ with ν and q. We will also assume that every $\nu^\dagger$ we refer to has $J(\nu^\dagger)$ finite.

Theorem 5.1 (Biased Sampling of ρ – Radon-Nikodym). *If $\rho^\varepsilon \ll \rho$ and*

$$\left\| 1 - \frac{d\rho^\varepsilon}{d\rho} \right\|_{L^\infty(\rho)} \leq \varepsilon, \tag{5.1}$$

then

$$\partial_t D^{p_t}(\mu^\dagger, \nu_t) < 0, \tag{5.2}$$

when

$$\|f - K\nu_t\|_{L^2(\rho^\varepsilon)} > (1+\varepsilon)\|f - K\mu^\dagger\|_{L^2(\rho)}. \tag{5.3}$$

Moreover, if $\mu^\dagger$ and $\nu^\dagger$ satisfy the source condition through $\phi \in L^2(\rho)$ and $\phi \in L^2(\rho^\varepsilon)$ respectively, then

$$\begin{aligned} D_J^{p_t}(\mu^\dagger, \mu_t) \leq \frac{1}{2t}\|\phi\|^2_{L^2(\rho)} + \frac{\varepsilon}{1+\varepsilon}\frac{1}{2t}\int_0^t \int_0^\tau \|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon)} ds d\tau \\ + (2\varepsilon+1)\frac{t}{4}\|f - K\mu^\dagger\|^2_{L^2(\rho)} + \frac{t}{4}\|f - K\nu^\dagger\|^2_{L^2(\rho^\varepsilon)} \end{aligned} \tag{5.4}$$

for almost every $t \geq 0$.

Theorem 5.2 (Biased Sampling of ρ – Wasserstein). *If $f \in \mathcal{C}^{0,1}(\mathrm{supp}(\rho - \rho^\varepsilon))$ and*

$$W_1(\rho, \rho^\varepsilon) \leq \varepsilon, \tag{5.5}$$

then

$$\partial_t D^{p_t}(\mu^\dagger, \nu_t) < 0, \tag{5.6}$$

when

$$\|f - K\nu_t\|^2_{L^2(\rho^\varepsilon)} > 2\varepsilon\|f - K\mu^\dagger\|^2_{\mathcal{C}^{(0,1)}(\mathrm{supp}(\rho - \rho^\varepsilon))} + \|f - K\mu^\dagger\|^2_{L^2(\rho)}. \tag{5.7}$$

Moreover, if $\mu^\dagger$ and $\nu^\dagger$ satisfy the source condition through $\phi \in L^2(\rho)$ and $\phi \in L^2(\rho^\varepsilon)$ respectively, then

$$\begin{aligned} D_J^{p_t}(\mu^\dagger, \nu_t) \leq \frac{1}{2t}\|\phi\|^2_{L^2(\rho)} + \varepsilon\frac{t}{2}\|f - K\mu^\dagger\|^2_{\mathcal{C}^{0,1}(\mathrm{supp}(\rho - \rho^\varepsilon))} \\ + \frac{\varepsilon}{t}\int_0^t \int_0^\tau \|K\nu_\tau - K\nu_s\|^2_{\mathcal{C}^{0,1}(\mathrm{supp}(\rho - \rho^\varepsilon))} ds d\tau + \frac{t}{4}\|K\nu^\dagger - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} \end{aligned} \tag{5.8}$$

for almost every $t \geq 0$.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Theorem 5.1 refers to the Radon-Nikodym derivative condition, whereas Theorem 5.2 refers to the Wasserstein-1 distance condition. To prove these theorems, we first consider a general disturbance with no particular conditions on the perturbation ρ^ε. Afterwards, we refine the statements from the general disturbance under the two mentioned conditions in Sections 5.1 and 5.2.

Lemma 5.1. *We have*

$$\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) \leq \frac{1}{4}\|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} \tag{5.9}$$

as well as

$$\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) \leq \frac{1}{4}\|K\nu^\dagger - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} \tag{5.10}$$

for almost every $t \geq 0$.

Proof. The first statement follows from

$$\begin{aligned}
&\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) \\
&\leq \langle \partial_t q_t \,|\, \nu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Lemma 3.2)} \\
&= \langle L_{\rho^\varepsilon}(f - K\nu_t) \,|\, \nu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Eq. (2.3))} \\
&= \langle f - K\nu_t \,|\, K(\nu_t - \mu^\dagger) \rangle_{L^2(\rho^\varepsilon)} && \text{(Lemma 2.1)} \\
&= \langle f - K\nu_t \,|\, K\nu_t - f + f - K\mu^\dagger \rangle_{L^2(\rho^\varepsilon)} \\
&= -\|f - K\nu_t\|^2_{L^2(\rho^\varepsilon)} + \langle f - K\mu^\dagger \,|\, f - K\nu_t \rangle_{L^2(\rho^\varepsilon)} \\
&\leq -\|f - K\nu_t\|^2_{L^2(\rho^\varepsilon)} + \|f - K\mu^\dagger\|_{L^2(\rho^\varepsilon)}\|f - K\nu_t\|_{L^2(\rho^\varepsilon)} && \text{(Cauchy-Schwarz)} \\
&\leq \frac{1}{4}\|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)}. && \text{(Young's product ineq.)}
\end{aligned}$$

The second statement follows from

$$\begin{aligned}
&\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) \\
&\leq \langle \partial_t q_t \,|\, \nu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Lemma 3.2)} \\
&= \langle L_{\rho^\varepsilon}(f - K\nu_t) \,|\, \nu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Eq. (2.3))} \\
&= \langle -L_{\rho^\varepsilon}(f - K\nu^\dagger) + L_{\rho^\varepsilon}(f - K\nu_t) \,|\, \nu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} && \text{(Proposition 2.5)} \\
&= \langle L_{\rho^\varepsilon}(K\nu^\dagger - K\nu_t) \,|\, \nu_t - \mu^\dagger \rangle_{\mathcal{M}(\Omega)} \\
&= \langle K\nu^\dagger - K\nu_t \,|\, K(\nu_t - \mu^\dagger) \rangle_{L^2(\rho^\varepsilon)} && \text{(Lemma 2.1)} \\
&= \langle K\nu^\dagger - K\nu_t \,|\, K\nu_t - K\nu^\dagger + K\nu^\dagger - K\mu^\dagger \rangle_{L^2(\rho^\varepsilon)} \\
&= -\|K\nu^\dagger - K\nu_t\|^2_{L^2(\rho^\varepsilon)} + \langle K\nu^\dagger - K\mu^\dagger \,|\, K\nu^\dagger - K\nu_t \rangle_{L^2(\rho^\varepsilon)} \\
&\leq -\|K\nu^\dagger - K\nu_t\|^2_{L^2(\rho^\varepsilon)} + \|K\nu^\dagger - K\mu^\dagger\|_{L^2(\rho^\varepsilon)}\|K\nu^\dagger - K\nu_t\|_{L^2(\rho^\varepsilon)} && \text{(Cauchy-Schwarz)} \\
&\leq \frac{1}{4}\|K\nu^\dagger - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)}. && \text{(Young's product ineq.)}
\end{aligned}$$

The proof is complete. □

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Proposition 5.1. *We have*

$$\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) < 0, \tag{5.11}$$

when

$$\|f - K\nu_t\|_{L^2(\rho^\varepsilon)} > \|f - K\mu^\dagger\|_{L^2(\rho^\varepsilon)}. \tag{5.12}$$

Proof. Recall from the proof of Lemma 5.1 that

$$\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) \leq -\|f - K\nu_t\|_{L^2(\rho^\varepsilon)}^2 + \|f - K\mu^\dagger\|_{L^2(\rho^\varepsilon)}\|f - K\nu_t\|_{L^2(\rho^\varepsilon)}. \tag{5.13}$$

Clearly, this is strictly negative when (5.12) is satisfied. □

Lemma 5.1 and Proposition 5.1 tell us, just like Lemma 4.1 for the noisy case, and as intuitively expected, that the flow will converge until the solution matches the residual. This, however, does not tell us how well it approximates the residual on ρ. We will refine this when we consider the more specific disturbances.

We will now provide an upper bound for the Bregman distance.

Proposition 5.2. *If $\mu^\dagger$ and $\nu^\dagger$ satisfy the source condition through $\phi \in L^2(\rho)$ and $\phi \in L^2(\rho^\varepsilon)$ respectively, then*

$$\begin{aligned} D_J^{p_t}(\mu^\dagger, \nu_t) \leq \frac{1}{2t}\|\phi\|_{L^2(\rho)}^2 + \frac{1}{2t}\int_0^t \int_0^\tau \|K\nu_\tau - K\nu_s\|_{L^2(\rho^\varepsilon - \rho)}^2 ds d\tau \\ + \frac{t}{4}\|f - K\mu^\dagger\|_{L^2(\rho^\varepsilon - \rho)}^2 + \frac{t}{8}\|K\nu^\dagger - K\mu^\dagger\|_{L^2(\rho^\varepsilon)}^2 \end{aligned} \tag{5.14}$$

for almost every $t \geq 0$.

Proof. Define

$$\partial_t e_t = K\mu^\dagger - K\nu_t, \quad e_0 = 0. \tag{5.15}$$

and

$$p^\dagger = L_\rho \phi. \tag{5.16}$$

With this we obtain

$$\begin{aligned} &\partial_t \left(\frac{1}{2}\|e_t - \phi\|_{L^2(\rho)}^2 \right) \\ &= \langle \partial_t e_t \,|\, e_t - \phi \rangle_{L^2(\rho)} \\ &= \langle K\mu^\dagger - K\nu_t \,|\, e_t - \phi \rangle_{L^2(\rho)} && \text{(Eq. (5.15))} \\ &= \langle L_\rho(e_t - \phi) \,|\, \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} && \text{(Lemma 2.1)} \\ &= \langle L_\rho(e_t - \phi) - q_t + q_t \,|\, \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} \\ &= \langle q_t - L_\rho\phi + L_\rho e_t - q_t \,|\, \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} \\ &= \langle q_t - p^\dagger \,|\, \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} + \langle L_\rho e_t - q_t \,|\, \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} && \text{(Eq. (5.16))} \\ &= -\left(D^{q_t}(\mu^\dagger, \nu_t) + D^{p^\dagger}(\nu_t, \mu^\dagger) \right) + \langle L_\rho e_t - q_t \,|\, \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)}. && \text{(Eq. (4.14))} \end{aligned}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

The rightmost term can be bounded by

$$
\begin{aligned}
&\langle L_\rho e_t - q_t \mid \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} \\
&= \int_0^t \langle \partial_s (L_\rho e_s - q_s) \mid \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} ds && \text{(Fund. th. of calc.)} \\
&= \int_0^t \langle L_\rho (K\mu^\dagger - K\nu_s) - L_{\rho^\varepsilon}(f - K\nu_s) \mid \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} ds && \text{(Eq. (5.15))} \\
&= \int_0^t \langle L_\rho (f - K\nu_s) - L_{\rho^\varepsilon}(f - K\nu_s) \mid \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} ds && \text{(Proposition 2.5)} \\
&= \int_0^t \langle L_{\rho - \rho^\varepsilon}(f - K\nu_s) \mid \mu^\dagger - \nu_t \rangle_{\mathcal{M}(\Omega)} ds \\
&= \int_0^t \langle f - K\nu_s \mid K\mu^\dagger - K\nu_t \rangle_{L^2(\rho - \rho^\varepsilon)} ds && \text{(Lemma 2.1)} \\
&= \int_0^t \langle f - K\nu_s \mid K\nu_t - K\mu^\dagger \rangle_{L^2(\rho^\varepsilon - \rho)} ds \\
&= \int_0^t \langle f - K\mu^\dagger - K\nu_t + K\mu^\dagger + K\nu_t - K\nu_s \mid K\nu_t - K\mu^\dagger \rangle_{L^2(\rho^\varepsilon - \rho)} ds \\
&= \int_0^t \langle f - K\mu^\dagger + K\nu_t - K\nu_s \mid K\nu_t - K\mu^\dagger \rangle_{L^2(\rho^\varepsilon - \rho)} \\
&\qquad - \|K\nu_t - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon - \rho)} ds \\
&\le \int_0^t \|f - K\mu^\dagger + K\nu_t - K\nu_s\|_{L^2(\rho^\varepsilon - \rho)} \|K\nu_t - K\mu^\dagger\|_{L^2(\rho^\varepsilon - \rho)} \\
&\qquad - \|K\nu_t - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon - \rho)} ds && \text{(Cauchy-Schwarz)} \\
&= \int_0^t \left(\|f - K\mu^\dagger\|_{L^2(\rho^\varepsilon - \rho)} + \|K\nu_t - K\nu_s\|_{L^2(\rho^\varepsilon - \rho)} \right) \\
&\qquad \times \|K\nu_t - K\mu^\dagger\|_{L^2(\rho^\varepsilon - \rho)} - \|K\nu_t - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon - \rho)} ds && \text{(Triangle ineq.)} \\
&\le \frac{1}{2} \int_0^t \|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon - \rho)} ds + \frac{1}{2} \int_0^t \|K\nu_t - K\nu_s\|^2_{L^2(\rho^\varepsilon - \rho)} ds && \text{(Young's prod. ineq.)} \\
&= \frac{t}{2} \|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon - \rho)} + \frac{1}{2} \int_0^t \|K\nu_t - K\nu_s\|^2_{L^2(\rho^\varepsilon - \rho)} ds.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\partial_t \left(\frac{1}{2} \|e_t - \phi\|^2_{L^2(\rho)} \right) + D^{p_t}(\mu^\dagger, \nu_t) \\
&\le \frac{t}{2} \|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon - \rho)} + \frac{1}{2} \int_0^t \|K\nu_t - K\nu_s\|^2_{L^2(\rho^\varepsilon - \rho)} ds. \qquad (5.17)
\end{aligned}
$$

Integrating from 0 to t gives

$$
\int_0^t D^{p_s}(\mu^\dagger, \nu_s) ds \le \frac{1}{2} \|\phi\|^2_{L^2(\rho)} + \frac{t^2}{4} \|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon - \rho)}
$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

$$+\frac{1}{2}\int_0^t\int_0^\tau \|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon-\rho)}dsd\tau. \tag{5.18}$$

Therefore, we obtain

$$\begin{aligned}
D^{p_t}(\mu^\dagger,\nu_t) &= \frac{1}{t}\int_0^t D^{p_t}(\mu^\dagger,\nu_t)ds \\
&\le \frac{1}{t}\int_0^t D^{p_s}(\mu^\dagger,\nu_s) + \int_s^t \partial_\tau D^{p_\tau}(\mu^\dagger,\nu_\tau)d\tau ds && \text{(Fund. th. of calc.)} \\
&\le \frac{1}{t}\int_0^t D^{p_s}(\mu^\dagger,\nu_s) + \frac{1}{4}\|K\nu^\dagger - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)}\int_s^t d\tau ds && \text{(Lemma 5.1)} \\
&= \frac{1}{t}\int_0^t D^{p_s}(\mu^\dagger,\nu_s) + \frac{1}{4}\|K\nu^\dagger - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)}(t-s)ds \\
&= \frac{1}{t}\int_0^t D^{p_s}(\mu^\dagger,\nu_s)ds + \frac{t}{8}\|K\nu^\dagger - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} \\
&\le \frac{1}{2t}\|\phi\|^2_{L^2(\rho)} + \frac{1}{2t}\int_0^t\int_0^\tau \|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon-\rho)}dsd\tau && \text{(Eq. (5.18))} \\
&\quad + \frac{t}{4}\|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon-\rho)} + \frac{t}{8}\|K\nu^\dagger - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)}.
\end{aligned}$$

The proof is complete. □

The bound of (5.14) in Proposition 5.2 is similar to that of (4.11) in Proposition 4.2. If ν_t remains constant for all t after some time $T \ge 0$, then

$$\frac{1}{2t}\int_0^t\int_0^\tau \|K\nu_\tau - K\nu_s\|_{L^2(\rho^\varepsilon-\rho)}dsd\tau = \mathcal{O}\left(1+\frac{1}{t}\right) \tag{5.19}$$

for all $t \ge T$. This implies that (5.14), just like (4.11), has a term that is inversely in time, a term constant in time and a term that is linearly increasing in time.

5.1 Radon-Nikodym

The first type of disturbances is expressed in terms of a bound on the Radon-Nikodym derivative. This allows for going from the norm using one measure to the norm using the other measure by adding a multiplicative constant.

For this subsection, we refine our definition of ρ^ε by assuming that ρ^ε is absolutely continuous with respect to ρ with

$$\left\|1 - \frac{d\rho^\varepsilon}{d\rho}\right\|_{L^\infty(\rho)} \le \varepsilon. \tag{5.20}$$

Lemma 5.2. *For all $g \in L^2(\rho)$,*

$$\|g\|^2_{L^2(\rho-\rho^\varepsilon)} \le \varepsilon\|g\|^2_{L^2(\rho)}, \tag{5.21}$$

$$\|g\|^2_{L^2(\rho^\varepsilon)} \le (1+\varepsilon)\|g\|^2_{L^2(\rho)}, \tag{5.22}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

and for all $g \in L^2(\rho^\varepsilon)$,

$$(1-\varepsilon)\|g\|^2_{L^2(\rho)} \leq \|g\|^2_{L^2(\rho^\varepsilon)}. \tag{5.23}$$

Proof. The first statement follows from

$$\begin{aligned}
\|g\|^2_{L^2(\rho-\rho^\varepsilon)} &= \int_{\mathcal{X}} g^2(x) d(\rho-\rho^\varepsilon)(x) \\
&= \int_{\mathcal{X}} g^2(x) \frac{d(\rho-\rho^\varepsilon)}{d\rho}(x) d\rho(x) \\
&\leq \left\| 1 - \frac{\mathrm{d}\rho^\varepsilon}{\mathrm{d}\rho} \right\|_{L^\infty(\rho)} \int_{\mathcal{X}} g^2(x) d\rho(x) \\
&\leq \varepsilon \|g\|^2_{L^2(\rho)}.
\end{aligned}$$

For the latter two observe that (5.20) means that

$$1-\varepsilon \leq \frac{\mathrm{d}\rho^\varepsilon}{\mathrm{d}\rho} \leq 1+\varepsilon \quad \rho \quad \text{a.e.}. \tag{5.24}$$

Hence,

$$\|g\|^2_{L^2(\rho^\varepsilon)} = \int_{\mathcal{X}} g^2(x) d\rho^\varepsilon(x) = \int_{\mathcal{X}} g^2(x) \frac{\mathrm{d}\rho^\varepsilon}{\mathrm{d}\rho}(x) d\rho(x) \leq (1+\varepsilon)\|g\|^2_{L^2(\rho)}$$

as well as

$$\|g\|^2_{L^2(\rho^\varepsilon)} = \int_{\mathcal{X}} g^2(x) d\rho^\varepsilon(x) = \int_{\mathcal{X}} g^2(x) \frac{\mathrm{d}\rho^\varepsilon}{\mathrm{d}\rho}(x) d\rho(x) \geq (1-\varepsilon)\|g\|^2_{L^2(\rho)}.$$

The proof is complete. □

Using the transformation rules of Lemma 5.2 we can provide conditions on when the rate of change of the Bregman distance is negative, similar to before.

Lemma 5.3. *We have*

$$\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) < 0 \tag{5.25}$$

for every $t \geq 0$, *when*

$$\|f - K\nu_t\|_{L^2(\rho^\varepsilon)} > (1+\varepsilon)\|f - K\mu^\dagger\|_{L^2(\rho)} \tag{5.26}$$

as well as when

$$\|f - K\nu_t\|_{L^2(\rho)} > \frac{1+\varepsilon}{1-\varepsilon}\|f - K\mu^\dagger\|_{L^2(\rho)} \tag{5.27}$$

and $\varepsilon < 1$.

Proof. Observe that

$$\begin{aligned}
\partial_t D_J^{q_t}(\mu^\dagger, \nu_t) &\leq \|f - K\mu^\dagger\|_{L^2(\rho^\varepsilon)} \|K\nu_t - f\|_{L^2(\rho^\varepsilon)} - \|K\nu_t - f\|^2_{L^2(\rho^\varepsilon)} && \text{(Eq. (5.13))} \\
&\leq (1+\varepsilon)\|f - K\mu^\dagger\|_{L^2(\rho)} \|K\nu_t - f\|_{L^2(\rho)} - \|K\nu_t - f\|^2_{L^2(\rho^\varepsilon)} && \text{(Eq. (5.22))} \\
&\leq (1+\varepsilon)\|f - K\mu^\dagger\|_{L^2(\rho)} \|K\nu_t - f\|_{L^2(\rho)} - (1-\varepsilon)\|K\nu_t - f\|^2_{L^2(\rho)}. && \text{(Eq. (5.23))}
\end{aligned}$$

Clearly, $\partial_t D_J^{q_t}(\mu^\dagger, \nu_t)$ is strictly negative when either (5.26) or (5.27) is satisfied. □

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

When comparing (5.26) with (4.9), we see that the sampling bias adds a multiplicative term based on ε. This is unlike the noisy case, where we got an additive term. Likewise, the upper bound for the Bregman distance also gets some multiplicative constants depending on ε.

Proposition 5.3. *If $\mu^\dagger$ and $\nu^\dagger$ satisfy the source condition through $\phi \in L^2(\rho)$ and $\phi \in L^2(\rho^\varepsilon)$ respectively, then*

$$D_J^{p_t}(\mu^\dagger, \mu_t) \leq \frac{1}{2t}\|\phi\|^2_{L^2(\rho)} + \frac{\varepsilon}{1+\varepsilon}\frac{1}{2t}\int_0^t \int_0^\tau \|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon)} ds d\tau \\ + (2\varepsilon+1)\frac{t}{4}\|f - K\mu^\dagger\|^2_{L^2(\rho)} + \frac{t}{4}\|f - K\nu^\dagger\|^2_{L^2(\rho^\varepsilon)} \tag{5.28}$$

for almost every $t \geq 0$.

Proof. From the transformation rules of Lemma 5.2 it follows that

$$\frac{t}{4}\|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon - \rho)} \leq \varepsilon\frac{t}{4}\|f - K\mu^\dagger\|^2_{L^2(\rho)} \tag{5.29}$$

as well as

$$\begin{aligned} &\|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon - \rho)} \\ &= \|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon)} - \|K\nu_\tau - K\nu_s\|^2_{L^2(\rho)} \\ &\leq \|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon)} - \frac{1}{1+\varepsilon}\|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon)} \\ &= \left(1 - \frac{1}{1+\varepsilon}\right)\|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon)} \\ &= \frac{\varepsilon}{1+\varepsilon}\|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon)}. \end{aligned} \tag{5.30}$$

Additionally,

$$\begin{aligned} &\|K\nu^\dagger - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} \\ &= \|K\nu^\dagger - f + f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} \\ &= \|K\nu^\dagger - f\|^2_{L^2(\rho^\varepsilon)} + \|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} \\ &\quad + 2\langle K\nu^\dagger - f \,|\, f - K\mu^\dagger\rangle_{L^2(\rho^\varepsilon)} \\ &\leq \|K\nu^\dagger - f\|^2_{L^2(\rho^\varepsilon)} + \|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} \\ &\quad + 2\|K\nu^\dagger - f\|_{L^2(\rho^\varepsilon)}\|f - K\mu^\dagger\|_{L^2(\rho^\varepsilon)} && \text{(Cauchy-Schwarz)} \\ &\leq 2\|K\nu^\dagger - f\|^2_{L^2(\rho^\varepsilon)} + 2\|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} && \text{(Young's product ineq.)} \\ &\leq 2\|K\nu^\dagger - f\|^2_{L^2(\rho^\varepsilon)} + 2(1+\varepsilon)\|f - K\mu^\dagger\|^2_{L^2(\rho)}. && \text{(Eq. (5.22))} \end{aligned} \tag{5.31}$$

Bounding (5.14) using (5.30), (5.29) and (5.31) gives the sought for expression. □

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Note that when we take the limit of $\varepsilon \to 0$ of (5.28), then we get

$$D_J^{p_t}(\mu^\dagger, \mu_t) \leq \frac{1}{2t}\|\phi\|^2_{L^2(\rho)} + \frac{t}{2}\|f - K\mu^\dagger\|^2_{L^2(\rho)}. \tag{5.32}$$

This shows that the bound for the Bregman distance in Proposition 5.3, unlike the bound in Proposition 5.2, is no longer tight in ε.

An interesting source of bias is when ρ^ε is a subsampling of ρ such that $\|f\|_{L^2(\rho^\varepsilon)}$ is a Monte Carlo estimator of $\|f\|_{L^2(\rho)}$. Clearly, $\rho^\varepsilon \ll \rho$ and ε is finite. This means that subsampling is a special case of Radon Nikodym bias and that we can use Proposition 5.3. At the same time, the fact that $\|f\|_{L^2(\rho^\varepsilon)}$ is a Monte Carlo estimator allows us to provide an alternative to (5.28).

Proposition 5.4. *Let $\rho \in \mathcal{P}_4(\mathcal{X})$ be a probability measure with bounded 4-th moment, ρ^ε be a subsampling of ρ with $m(\varepsilon) \in \mathbb{N}$ samples, $\delta > 0$, and $f \in L^2(\rho) \cap L^4(\rho)$. If $\mu^\dagger$ and $\nu^\dagger$ satisfy the source condition through $\phi \in L^2(\rho)$ and $\phi \in L^2(\rho^\varepsilon)$ respectively, then*

$$\begin{aligned} D_J^{p_t}(\mu^\dagger, \nu_t) \leq \frac{1}{2t}\|\phi\|^2_{L^2(\rho)} + \frac{1}{2t\sqrt{m(\varepsilon)\delta}} \int_0^t \int_0^\tau \|K\nu_\tau - K\nu_s\|^2_{L^4(\rho)} ds d\tau \\ + \frac{t}{4\sqrt{m(\varepsilon)\delta}}\|f - K\mu^\dagger\|^2_{L^4(\rho)} + \frac{t}{8}\|K\nu^\dagger - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} \end{aligned} \tag{5.33}$$

for almost every $t \geq 0$ with probability at least $1 - \delta$.

Proof. Since ρ has bounded 4-th moment, we get by Proposition 1.2 that $K\mu \in L^4(\rho)$ for all $\mu \in \mathcal{M}(\Omega)$.

From Chebychev's inequality it follows that

$$\begin{aligned} &\left| \|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon - \rho)} \right|^2 \\ &= \left| \int_{\mathcal{X}} |K\nu_\tau(x) - K\nu_s(x)|^2 d\rho^\varepsilon(x) - \int_{\mathcal{X}} |K\nu_\tau(x) - K\nu_s(x)|^2 d\rho(x) \right|^2 \\ &\leq \frac{\int_{\mathcal{X}} |K\nu_\tau(x) - K\nu_s(x)|^4 d\rho(x) - \left(\int_{\mathcal{X}} |K\nu_\tau(x) - K\nu_s(x)|^2 d\rho(x) \right)^2}{m(\varepsilon)\delta} \\ &= \frac{\|K\nu_\tau - K\nu_s\|^4_{L^4(\rho)} - \|K\nu_\tau - K\nu_s\|^4_{L^2(\rho)}}{m(\varepsilon)\delta} \\ &\leq \frac{\|K\nu_\tau - K\nu_s\|^4_{L^4(\rho)}}{m(\varepsilon)\delta} \end{aligned} \tag{5.34}$$

with probability at least $1 - \delta$. Taking the square root on both sides gives

$$\|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon - \rho)} \leq \frac{\|K\nu_\tau - K\nu_s\|^2_{L^4(\rho)}}{\sqrt{m(\varepsilon)\delta}}. \tag{5.35}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Similarly,

$$\|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon - \rho)} \leq \frac{\|f - K\mu^\dagger\|^2_{L^4(\rho)}}{\sqrt{m(\varepsilon)\delta}}. \tag{5.36}$$

Substitution of (5.35) and (5.36) into (5.14) gives (5.33). □

Note that when we take the limit of $m(\varepsilon) \to \infty$ of (5.33), then we get

$$D_J^{p_t}(\mu^\dagger, \mu_t) \leq \frac{1}{2t}\|\phi\|^2_{L^2(\rho)}. \tag{5.37}$$

This shows that the bound for the Bregman distance in Proposition 5.4, like the bound in Proposition 5.2, is tight in ε.

5.2 Wasserstein

The second type of disturbances is expressed in terms of a bound on the Wasserstein metric. This allows for going from the norm using one measure to the norm using the other measure by using the duality between Wasserstein and the Lipschitz continuous function with Lipschitz constant at most 1.

For this subsection, we refine our definition of ρ^ε by assuming that the Wasserstein-1 distance between ρ^ε and ρ is bounded through ε, i.e.

$$W_1(\rho^\varepsilon, \rho) \leq \varepsilon. \tag{5.38}$$

We also assume that $f \in \mathcal{C}^{0,1}(\mathrm{supp}(\rho - \rho^\varepsilon))$.

Lemma 5.4. *For all $g \in \mathcal{C}^{0,1}(\mathrm{supp}(\rho - \rho^\varepsilon))$,*

$$\|g\|^2_{L^2(\rho^\varepsilon - \rho)} \leq 2\|g\|^2_{\mathcal{C}^{0,1}(\mathrm{supp}(\rho^\varepsilon - \rho))}\varepsilon. \tag{5.39}$$

Proof. Recall that

$$W_1(\rho^\varepsilon, \rho) = \sup_{\substack{h \in \mathcal{C}^{0,1}(\mathcal{X}) \\ Lip(h) \leq 1}} \langle h \,|\, \rho^\varepsilon - \rho\rangle_{\mathcal{M}(\mathcal{X})}.$$

Since for all $g \in \mathcal{C}^{0,1}(\mathrm{supp}(\rho^\varepsilon - \rho))$,

$$\mathrm{Lip}\left(\frac{g}{\mathrm{Lip}(g)}\right) \leq 1, \tag{5.40}$$

we obtain

$$\begin{aligned}\langle g \,|\, \rho^\varepsilon - \rho\rangle_{\mathcal{M}(\mathcal{X})} &= \mathrm{Lip}(g)\left\langle \frac{g}{\mathrm{Lip}(g)} \,\middle|\, \rho^\varepsilon - \rho \right\rangle_{\mathcal{M}(\mathcal{X})} \\ &\leq \mathrm{Lip}(g) W_1(\rho^\varepsilon, \rho) \leq \mathrm{Lip}(g)\varepsilon,\end{aligned} \tag{5.41}$$

where we used (5.38). Furthermore,

$$\mathrm{Lip}(|g|^2) \leq 2\|g\|^2_{\mathcal{C}^{0,1}(\mathrm{supp}(\rho^\varepsilon - \rho))} < \infty,$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

since

$$\begin{aligned}\big||g(x)|^2 - |g(y)|^2\big| &= |g(x) - g(y)|\,\big||g(x)| + |g(y)|\big| \\ &\leq 2\|g\|_{\mathcal{C}(\mathrm{supp}(\rho^\varepsilon - \rho))}\mathrm{Lip}(g)\|x - y\|_{\ell^\infty} \end{aligned} \tag{5.42}$$

for all $x, y \in \mathrm{supp}(\rho^\varepsilon - \rho)$. Hence,

$$\begin{aligned}\|g\|^2_{L^2(\rho^\varepsilon - \rho)} &= \langle |g|^2 \,|\, \rho^\varepsilon - \rho\rangle_{\mathcal{M}(\mathcal{X})} \\ &\leq \mathrm{Lip}(|g|^2)\varepsilon && \text{(Eq. (5.41))} \\ &\leq 2\|g\|^2_{\mathcal{C}^{0,1}(\mathcal{X})}\varepsilon && \text{(Eq. (5.42))}\end{aligned}$$

for all $g \in \mathcal{C}^{0,1}(\mathrm{supp}(\rho^\varepsilon - \rho))$. □

Proposition 5.5. *We have*

$$\partial_t D^{q_t}(\mu^\dagger, \nu_t) < 0, \tag{5.43}$$

when

$$\|K\nu_t - f\|^2_{L^2(\rho^\varepsilon)} > 2\varepsilon\|f - K\mu^\dagger\|^2_{\mathcal{C}^{0,1}(\mathrm{supp}(\rho - \rho^\varepsilon))} + \|f - K\mu^\dagger\|^2_{L^2(\rho)}. \tag{5.44}$$

Proof. $f - K\mu^\dagger$ is a sum of two Lipschitz functions on $\mathrm{supp}(\rho - \rho^\varepsilon)$; f by assumption and $K\mu^\dagger$ by Proposition 1.1. Thus, $f - K\mu^\dagger$ is Lipschitz on $\mathrm{supp}(\rho - \rho^\varepsilon)$. From Lemma 5.4 we obtain that

$$\|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)} \leq 2\varepsilon\|f - K\mu^\dagger\|^2_{\mathcal{C}^{0,1}(\mathrm{supp}(\rho - \rho^\varepsilon))} + \|f - K\mu^\dagger\|^2_{L^2(\rho)}. \tag{5.45}$$

Hence,

$$\begin{aligned}\partial_t D^{q_t}_J(\mu^\dagger, \nu_t) &\leq \|f - K\mu^\dagger\|_{L^2(\rho^\varepsilon)}\|K\nu_t - f\|_{L^2(\rho^\varepsilon)} - \|K\nu_t - f\|^2_{L^2(\rho^\varepsilon)} && \text{(Eq. (5.13))} \\ &\leq \sqrt{2\varepsilon\|f - K\mu^\dagger\|^2_{\mathcal{C}^{0,1}(\mathrm{supp}(\rho - \rho^\varepsilon))} + \|f - K\mu^\dagger\|^2_{L^2(\rho)}} \\ &\quad \times \|K\nu_t - f\|_{L^2(\rho^\varepsilon)} - \|K\nu_t - f\|^2_{L^2(\rho^\varepsilon)}. && \text{(Eq. (5.45))}\end{aligned}$$

Clearly, this is strictly negative when (5.44) is satisfied. □

When comparing (5.26) with (4.9), we see that the sampling bias adds an additive term based on ε. This is like the noisy case, but unlike when the sampling bias was given in terms of the Radon-Nikodym derivative.

Proposition 5.6. *If $\mu^\dagger$ and $\nu^\dagger$ satisfy the source condition through $\phi \in L^2(\rho)$ and $\phi \in L^2(\rho^\varepsilon)$ respectively, then*

$$\begin{aligned}D^{p_t}_J(\mu^\dagger, \nu_t) \leq \frac{1}{2t}\|\phi\|^2_{L^2(\rho)} + \varepsilon\frac{t}{2}\|f - K\mu^\dagger\|^2_{\mathcal{C}^{0,1}(\mathrm{supp}(\rho - \rho^\varepsilon))} \\ + \frac{\varepsilon}{t}\int_0^t\int_0^\tau \|K\nu_\tau - K\nu_s\|^2_{\mathcal{C}^{0,1}(\mathrm{supp}(\rho - \rho^\varepsilon))} ds d\tau + \frac{t}{8}\|K\nu^\dagger - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon)}\end{aligned} \tag{5.46}$$

for almost every $t \geq 0$.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

Proof. Recall from the proof of Proposition 5.5 that $f - K\mu^\dagger \in \mathcal{C}^{0,1}(\operatorname{supp}(\rho - \rho^\varepsilon))$. Eq. (5.45) can be rewritten as

$$\|f - K\mu^\dagger\|^2_{L^2(\rho^\varepsilon - \rho)} \leq 2\varepsilon \|f - K\mu^\dagger\|^2_{\mathcal{C}^{0,1}(\operatorname{supp}(\rho - \rho^\varepsilon))}. \tag{5.47}$$

Similarly, $K\nu_\tau - K\nu_s \in \mathcal{C}^{0,1}(\operatorname{supp}(\rho - \rho^\varepsilon))$ by Proposition 1.2. Hence,

$$\|K\nu_\tau - K\nu_s\|^2_{L^2(\rho^\varepsilon - \rho)} \leq 2\varepsilon \|K\nu_\tau - K\nu_s\|^2_{\mathcal{C}^{0,1}(\operatorname{supp}(\rho - \rho^\varepsilon))}. \tag{5.48}$$

Bounding (5.14) using (5.47) and (5.48) gives the sought for expression. □

Note that when we take the limit of $\varepsilon \to 0$ of (5.28), then we get

$$D_J^{p_t}(\mu^\dagger, \mu_t) \leq \frac{1}{2t} \|\phi\|^2_{L^2(\rho)}. \tag{5.49}$$

This shows that the bound for the Bregman distance in Proposition 5.6, like the bound in Proposition 5.2, is tight in ε.

6 Parameter space discretization

One issue with the inverse scale space of (1.11) is that p_t is defined on Ω. To ensure that $p_t \in \partial J(\mu_t)$ we need to have full knowledge of p_t. This cannot be implemented. Hence, Ω needs to be discretized. In this section, we study a particular discretization based on the Voronoi tessellation. In Section 6.1, we show, for a given sequence of Voronoi tessellations with mild assumptions, that the inverse scale space flow on these tessellations converges to the full flow for $N \to \infty$. In Section 6.2, we show the rate of convergence for the flow with fixed N to the optimal solution. Combined, these sections prove Theorem 6.1.

Given a set $\omega^N \subseteq \Omega$ with $|\omega^N| = N$, a Voronoi tessellation divides Ω into N subsets

$$\Omega_n^N = \left\{ w \in \Omega \;\middle|\; \forall\, m \in \{1, \dots, N\} : \; \left|w - \omega_n^N\right| \leq \left|w - \omega_m^N\right| \right\} \tag{6.1}$$

such that

$$\Omega = \bigcup_{n=1}^{N} \Omega_n^N. \tag{6.2}$$

We consider sequences of sets $\{\omega^N\}_{N=1}^\infty$ with $\omega^N \subseteq \Omega$, $|\omega^N| = N$ and

$$\lim_{N \to \infty} \max_n \operatorname{diam}\left(\Omega_n^N\right) = 0.$$

For this section, we will keep referring to the solution over Ω with μ and p whilst we will refer to the solution over ω^N with ν and q. With $\nu^\dagger$ we denote a minimizer of $\mathcal{R}_f$ with $J(\nu^\dagger) < \infty$ over the measures supported on ω^N. We will make use of the Lagrangian

$$F : \mathcal{M}(\Omega) \;\to\; [0, \infty), \quad \mu \;\mapsto\; J(\mu) + \lambda \mathcal{R}_f(\mu) \tag{6.3}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

of (2.3) and its restriction to ω^N

$$F_N\mu = \begin{cases} F\mu, & \operatorname{supp}(\mu) \subseteq \omega^N, \\ \infty, & \text{otherwise} \end{cases} \tag{6.4}$$

in the proofs. We will also assume that Ω is compact.

Theorem 6.1. *The minimizers of the sequence $\{F_N\}_{N=1}^{\infty}$ converge in weak* to the minimizer of F. Moreover,*

$$\begin{aligned} \|K\nu_t - f\|^2_{L^2(\rho)} \leq\ & 2\|K\mu^\dagger - f\|^2_{L^2(\rho)} \\ & + 2\mathrm{Lip}(\sigma)^2 \Big(\max_n \operatorname{diam}\left(\Omega_n^N\right)\Big)^2 \|\mu^\dagger\|^2 + 2\frac{J(\nu^\dagger)}{t} \end{aligned} \tag{6.5}$$

for almost every $t \geq 0$.

The first part of the theorem follows from Section 6.1 and the second part from Section 6.2.

6.1 Convergence of the discrete flow to the full flow

Both the discrete flow and full flow are well-defined flows, so what remains to show is that the solutions to the discrete flow for increasing N converge to the solution for the full flow. To prove this, we will show that the minimizers associated to the Lagrangians of the discrete flows F_N converge in weak* to the minimizer of F. This can be concluded from the fundamental theorem of Γ-convergence [7]. The requirements for the fundamental theorem are that F_N satisfies the lim inf property, that there exists a Γ-realizing sequence and that the family $(F_N)_N$ is equicoercive. The three propositions at the end of this subsection show that these requirements hold. These propositions rely on some properties of F that carry over to F_N. We will prove those first.

Lemma 6.1. *F is proper, convex, weak* lower semi-continuous and coercive.*

Proof. F is proper, since $0 \in \operatorname{dom}(F)$.

Since V is continuous, J is convex. Since K is a bounded, linear (and thus continuous) operator and the square of the $L^2(\rho)$ norm is convex, $\mathcal{R}_f$ is convex. Since F is a sum of two convex functions, F is convex.

Let (μ_n) be a sequence of measures and $\mu_n, \mu \in \mathcal{M}(\Omega)$ such that $\mu_n \xrightarrow{w^\star} \mu$. Then for all $\phi \in L^2(\rho)$,

$$\lim_{n\to\infty} \langle K\mu_n \,|\, \phi\rangle_{L^2(\rho)} = \lim_{n\to\infty} \langle L_\rho\phi \,|\, \mu_n\rangle_{\mathcal{M}(\Omega)} = \langle L_\rho\phi \,|\, \mu\rangle_{\mathcal{M}(\Omega)} = \langle K\mu \,|\, \phi\rangle_{L^2(\rho)}. \tag{6.6}$$

This shows that $K\mu_n \xrightarrow{L^2(\rho)} K\mu$. Since

$$w \ \mapsto\ \frac{1}{2}\|w - f\|^2_{L^2(\rho)} \tag{6.7}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

is continuous and convex, it is sequentially weak lower-semicontinuous. The combination implies that $\mathcal{R}_f$ is sequentially weak* lower-semicontinuous. Since J is continuous, it is weak* lower-semicontinuous. This implies that F is weak* lower-semicontinuous.

F is coercive if and only if

$$\lim_{\|\mu\|_{\mathcal{M}(\Omega)}\to\infty} F(\mu) = \infty. \tag{6.8}$$

For measures $\mu \notin N(K)$ outside the kernel of K we have that $\mathcal{R}_f(\mu) \to \infty$ as $\|\mu\|_{\mathcal{M}(\Omega)} \to \infty$. Since J is non-negative, F will grow without bound for those measures too. What remains is the measures $\mu \in N(K)$ inside the kernel of K. For these measures $\mathcal{R}_f(\mu)$ is constant, but the conditions on V imply that J will grow without bound. Hence, F is coercive. □

Now, we can prove the three properties needed for the sequence of F_N's.

Proposition 6.1 (Liminf Property)**.** *For all $\mu \in \mathcal{M}(\Omega)$ and every sequence (μ_n) such that $\mu_n \xrightarrow{w^*} \mu$, we have*

$$\liminf_{\mu_n\to\infty} F_n(\mu_n) \geq F(\mu). \tag{6.9}$$

Proof. From construction of F_n it follows that

$$F_n(\mu) \geq F(\mu). \tag{6.10}$$

Hence, combined with the lower semi-continuity of F proven in Lemma 6.1, we obtain

$$\liminf_{\mu_n\to\infty} F_n(\mu_n) \geq \liminf_{\mu_n\to\infty} F(\mu_n) \geq F(\mu). \tag{6.11}$$

The proof is complete. □

Proposition 6.2 (Γ-Realizing Sequence)**.** *Let $\mu \in \mathcal{M}(\Omega)$ and define a sequence of measures $\mu_N \in \mathcal{M}(\Omega)$ by*

$$\mu_N = \sum_{n=1}^{N} \mu\big(\Omega_n^N\big)\delta_{\omega_n^N}. \tag{6.12}$$

We have $\mu_N \xrightarrow{w^*} \mu$ as well as

$$\lim_{N\to\infty} F_N(\mu_N) = F(\mu). \tag{6.13}$$

Proof. Recall that $\mathcal{M}(\Omega)$ is dual to $C(\Omega)$, so the weak* convergence is defined in terms of $g \in \mathcal{C}(\Omega)$. Since Ω is compact, g is absolutely continuous. Recall that this implies that

$$\begin{aligned} &\forall\, \varepsilon > 0 \quad \exists\, \delta > 0 \quad \forall\, (a,b),(c,d) \in \Omega : \\ &\quad \|(a,b) - (c,d)\| < \delta \quad \Longrightarrow \quad |g(a,b) - g(c,d)| < \varepsilon. \end{aligned} \tag{6.14}$$

Since the diameter of the Voronoi cells vanishes as N goes to infinity, there must be an $\tilde{N}$ such that for all $N > \tilde{N}$ and $n \in \{1,\dots,N\}$, we have that $\|(a,b) - (a_n^N, b_n^N)\| < \delta$ for all $(a,b) \in \Omega_n^N$. Hence, for all $g \in \mathcal{C}(\Omega)$ and all $\varepsilon > 0$,

$$\lim_{N\to\infty} \left| \int_\Omega g(a,b) d(\mu - \mu_N)(a,b) \right|$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48

OPEN ACCESS

$$
\begin{aligned}
&= \lim_{N\to\infty} \left| \int_\Omega g(a,b) d\left(\mu - \sum_{n=1}^{N} \mu(\Omega_n^N) \delta_{\omega_n^N} \right)(a,b) \right| \\
&= \lim_{N\to\infty} \left| \int_\Omega g(a,b) d\mu(a,b) - \sum_{n=1}^{N} g(a_n^N, b_n^N) \mu(\Omega_n^N) \right| \\
&= \lim_{N\to\infty} \left| \int_\Omega g(a,b) d\mu(a,b) - \sum_{n=1}^{N} \int_{\Omega_n^N} g(a_n^N, b_n^N) d\mu(a,b) \right| \\
&= \lim_{N\to\infty} \left| \sum_{n=1}^{N} \int_{\Omega_n^N} g(a,b) d\mu(a,b) - \sum_{n=1}^{N} \int_{\Omega_n^N} g(a_n^N, b_n^N) d\mu(a,b) \right| \\
&= \lim_{N\to\infty} \left| \sum_{n=1}^{N} \int_{\Omega_n^N} \left(g(a,b) - g(a_n^N, b_n^N) \right) d\mu(a,b) \right| \\
&\leq \lim_{N\to\infty} \sum_{n=1}^{N} \int_{\Omega_n^N} \left\| g(a,b) - g(a_n^N, b_n^N) \right\| d|\mu|(a,b) \\
&< \lim_{N\to\infty} \sum_{n=1}^{N} \int_{\Omega_n^N} \varepsilon d|\mu|(a,b) \\
&= \varepsilon \|\mu\|_{\mathcal{M}(\Omega)}.
\end{aligned}
$$

Since ε was arbitrary, we must have that

$$
\lim_{N\to\infty} \int_\Omega g(a,b) d(\mu - \mu_N)(a,b) = 0. \tag{6.15}
$$

This shows that $\mu_N \xrightarrow{w^*} \mu$, and by construction of μ_N we have $F_N(\mu_N) = F(\mu_N)$. Furthermore, we showed in Lemma 6.1 that F was weak* lower semi-continuous. If fact, by similar arguments, it is sequentially weak* continuous. Hence, it follows that

$$
\lim_{N\to\infty} F_N(\mu_N) = \lim_{N\to\infty} F(\mu_N) = F(\mu). \tag{6.16}
$$

The proof is complete. □

Proposition 6.3 (Equicoercivity)**.** *The family $(F_N)_N$ is equicoercive.*

Proof. The family $(F_N)_N$ is equicoercive if and only if every member of the family is coercive. In Lemma 6.1 it was proven that F is coercive. Hence, by construction of F_N

$$
\lim_{\|\mu\|_{\mathcal{M}(\Omega)}\to\infty} F_N(\mu) \geq \lim_{\|\mu\|_{\mathcal{M}(\Omega)}\to\infty} F(\mu) = \infty. \tag{6.17}
$$

This means that F_N is coercive. Since N was arbitrary, it holds for all members F_N of the family $(F_N)_N$. □

We have now shown that the requirements for the fundamental theorem of Γ-convergence hold, which implies that the sequence of minimizers of F_N converges in weak* to the minimizer of F.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

6.2 Convergence error for the discrete flow

In the previous section, we showed that the discrete flow converges to the full flow. In this section, we will fix N and show the convergence rates of the discrete flow to the optimal solution. We will first show the generic bound, also shown in Theorem 6.1. Afterward, we will look at a special case.

Observe that the finite ω^N satisfies the required properties for a proper inverse scale space flow. The following proposition shows the generic bound.

Proposition 6.4. *We have*

$$\begin{aligned}\|Kv_t - f\|_{L^2(\rho)}^2 \leq 2\|K\mu^\dagger - f\|_{L^2(\rho)}^2 + 2\frac{J(v^\dagger)}{t} \\ + 2\|\mu^\dagger\|_{\mathcal{M}(\Omega)}^2 \Big(\max_n \operatorname{diam}\left(\Omega_n^N\right)\Big)^2 \operatorname{Lip}(\sigma)^2 \int_{\mathcal{X}} \max(1, \|x\|)^2 d\rho(x)\end{aligned} \tag{6.18}$$

for almost every $t \geq 0$.

Proof. From Proposition 3.3 it follows that

$$\|Kv_t - f\|_{L^2(\rho)}^2 \leq \|Kv^\dagger - f\|_{L^2(\rho)}^2 + 2\frac{J(v^\dagger)}{t}. \tag{6.19}$$

Since $v^\dagger$ is a minimizer of $\mathcal{R}_f$ over ω^N, we have for the measure

$$\mu_N = \sum_{n=1}^{N} \mu^\dagger\left(\Omega_n^N\right)\delta_{\omega_n^N}, \tag{6.20}$$

that

$$\begin{aligned}\|Kv^\dagger - f\|_{L^2(\rho)} &\leq \|K\mu_N - f\|_{L^2(\rho)} \\ &\leq \|K\mu_N - K\mu^\dagger\|_{L^2(\rho)} + \|K\mu^\dagger - f\|_{L^2(\rho)},\end{aligned} \tag{6.21}$$

and thus by Young's inequality for products with $p = q = 2$,

$$\|Kv^\dagger - f\|_{L^2(\rho)}^2 \leq 2\|K\mu_N - K\mu^\dagger\|_{L^2(\rho)}^2 + 2\|K\mu^\dagger - f\|_{L^2(\rho)}^2. \tag{6.22}$$

We observe that by a similar argument as in the proof of Proposition 6.2 that

$$\begin{aligned}&\|K\mu_N - K\mu^\dagger\|_{L^2(\rho)}^2 \\ &= \int_{\mathcal{X}} \left| \int_\Omega \sigma(a^\intercal x + b) d(\mu_n - \mu^\dagger)(a,b) \right|^2 d\rho(x) \\ &\leq \int_{\mathcal{X}} \left(\sum_{n=1}^{N} \int_{\Omega_n^N} \left\| \sigma(a^\intercal x + b) - \sigma\left(\left(a_n^N\right)^\intercal x + b_n^N\right) \right\| d|\mu^\dagger|(a,b) \right)^2 d\rho(x)\end{aligned}$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

$$
\begin{aligned}
&\leq \int_{\mathcal{X}} \left(\sum_{n=1}^{N} \int_{\Omega_n^N} \mathrm{Lip}(\sigma) \left\| (a^\intercal x + b) - \left((a_n^N)^\intercal x + b_n^N \right) \right\| d|\mu^\dagger|(a,b) \right)^2 d\rho(x) \\
&\leq \int_{\mathcal{X}} \left(\sum_{n=1}^{N} \int_{\Omega_n^N} \mathrm{Lip}(\sigma) \left(\left\| a - a_n^N \right\| \|x\| + \left| b - b_n^N \right| \right) d|\mu^\dagger|(a,b) \right)^2 d\rho(x) \\
&\leq \int_{\mathcal{X}} \max(1, \|x\|)^2 \left(\sum_{n=1}^{N} \int_{\Omega_n^N} \mathrm{Lip}(\sigma) \left(\left\| a - a_n^N \right\| + \left| b - b_n^N \right| \right) d|\mu^\dagger|(a,b) \right)^2 d\rho(x) \\
&= \mathrm{Lip}(\sigma)^2 \int_{\mathcal{X}} \max(1, \|x\|)^2 \left(\sum_{n=1}^{N} \int_{\Omega_n^N} \mathrm{diam}\left(\Omega_n^N\right) d|\mu^\dagger|(a,b) \right)^2 d\rho(x) \\
&\leq \|\mu^\dagger\|_{\mathcal{M}(\Omega)}^2 \left(\max_n \mathrm{diam}\left(\Omega_n^N\right) \right)^2 \mathrm{Lip}(\sigma)^2 \int_{\mathcal{X}} \max(1, \|x\|)^2 d\rho(x).
\end{aligned}
$$

Substituting this into (6.22) and the resulting expression into (6.19) gives (6.18). □

In [23], it was shown that a Voronoi cell's radius decreases with a rate of $\mathcal{O}(N^{-1/d})$ when the points in ω^N are i.i.d. sampled from an absolutely continuous probability measure over Ω. We can use the direct approximation theorem of Barron spaces to achieve a better rate [26, Theorem 3.8].

Proposition 6.5. *Let $N \in \mathbb{N}$. Denote with M_f the set of all measures μ_N of N atoms that satisfy the bounds*

$$
\|K\mu_N - K\mu^\dagger\|_{L^2(\rho)}^2 \leq \frac{J(\mu^\dagger)^2}{N} \mathrm{Lip}(\sigma)^2 \int_{\mathcal{X}} \max(1 + \|x\|)^2 d\rho(x), \tag{6.23}
$$

and choose ω^N such that M_f is non-empty. Then,

$$
\begin{aligned}
\|K\nu_t - f\|_{L^2(\rho)}^2 \leq\ & 3\|K\mu^\dagger - f\|_{L^2(\rho)}^2 + 2\frac{J(\nu^\dagger)}{t} \\
&+ 3\frac{J(\mu^\dagger)^2}{N} \mathrm{Lip}(\sigma)^2 \int_{\mathcal{X}} \max(1, \|x\|)^2 d\rho(x) \\
&+ 3 \inf_{\mu_N \in M_f} \|K\nu^\dagger - K\mu_N\|_{L^2(\rho)}^2.
\end{aligned} \tag{6.24}
$$

Proof. $K\mu^\dagger \in \mathcal{B}$, so by [24, Theorem 4] there exists a suitable choice for ω^N. Let $\mu_N \in M_f$. Observe that

$$
\begin{aligned}
\|K\nu_t - f\|_{L^2(\rho)}^2 &\leq \|K\nu^\dagger - f\|_{L^2(\rho)}^2 + 2\frac{J(\nu^\dagger)}{t} && \text{(Proposition 3.3)} \\
&\leq 3\|K\nu^\dagger - K\mu_N\|_{L^2(\rho)}^2 + 3\|K\mu^\dagger - K\mu_N\|_{L^2(\rho)}^2 \\
&\quad + 3\|K\mu^\dagger - f\|_{L^2(\rho)}^2 + 2\frac{J(\nu^\dagger)}{t} && \text{(Triangle ineq., Young's)} \\
&\leq 3\|K\nu^\dagger - K\mu_N\|_{L^2(\rho)}^2
\end{aligned}
$$

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

$$+ 3\frac{J(\mu^\dagger)^2}{N}\mathrm{Lip}(\sigma)^2 \int_{\mathcal{X}} \max(1, \|x\|)^2 d\rho(x)$$
$$+ 3\|K\mu^\dagger - f\|^2_{L^2(\rho)} + 2\frac{J(\nu^\dagger)}{t}. \qquad \text{(Eq. (6.23))}$$

Taking the infimum over $\mu_N \in M_f$ gives (6.24). □

7 Discussion

In this work, we have studied the convergence and error analysis of finding the best measure μ such that the Barron function $K\mu$ is close to f using the inverse scale space flow. After having established the existence and regularity of the solution, we considered the ideal, noisy, biased, and discretized cases. For each of these cases, we analysed the evolution of the Bregman divergence with respect to the optimal solution $D^{p_t}(\mu^\dagger, \nu_t)$ and the L^2 loss $\mathcal{R}_f(\mu_t)$.

In the ideal case, we got monotonic and linear evolution to the optimal solution. In the noisy case, we still got monotonic and linear evolution to the optimal solution but only up to an error level determined by the noise level δ. These results agree with the known results for inverse scale spaces.

In the novel case of biased sampling,

$$D^{p_t}(\mu^\dagger, \nu_t) \leq \mathcal{O}\left(1 + \frac{1}{t} + t\right)$$

with the suppressed factors in the big $\mathcal{O}$ notation depending on ε. When we work with noisy measurements, $D^{p_t}(\mu^\dagger, \nu_t)$ has a similar upper bound but depending on δ. In that setting, the smallest upper bound for $D^{p_t}(\mu^\dagger, \nu_t)$ is attained for $t(\delta) = \mathcal{O}(\delta^{-1})$. When dealing with biased sampling, this smallest upper bound is attained for

$$t(\varepsilon) = \mathcal{O}\left(\frac{\sqrt{1+\varepsilon}}{\sqrt{1+\varepsilon+\varepsilon^2}}\right), \quad t(\varepsilon) = \mathcal{O}\left(\frac{\sqrt{1+\varepsilon}}{\sqrt{\varepsilon}}\right)$$

for a Radon-Nikodym and a Wasserstein perturbation respectfully. However, whilst in many cases it is straightforward to provide an estimate for δ, it is not the case for ε.

A second issue with the upper bounds for $D^{p_t}(\mu^\dagger, \nu_t)$ is that we typically do not know $f, \phi, \mu^\dagger, \nu^\dagger$ or ρ. What we do know is $K\nu_t$ on $\mathrm{supp}(\rho^\varepsilon)$. This means the bound in Proposition 5.3 has more terms that can be explicitly computed than the bounds in Proposition 5.2, Proposition 5.4 or Proposition 5.6. That makes Proposition 5.3 arguably the most useful proposition.

When the parameter space Ω is discretized, we have shown that we still have a proper inverse scale space flow. In this setting, we get an additional additive factor depending on N in convergence. When we don't make any additional assumptions on ω^N, this additional factor is of the form $\mathcal{O}(N^{-1/d})$. This $1/d$ factor shows that the discretization method suffers from the curse of dimensionality, meaning that the method performs poorly when working with high dimension. Although we show that an $\mathcal{O}(N^{-1/2})$ can be attained in

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

theory, it is unclear how to find the required N points without solving a different sparse minimization problem first.

When applied in practice, the true data distribution needs to be sampled. This introduces an error that can either be bounded by a Monte-Carlo rate or using one of the approaches outlined in Section 5. Whilst the sampling introduces an error, it allows for potential improvements to the iterative scheme. When the data distribution is concentrated on finitely many atoms, we get for the chosen regularizer J that there exists a minimizing measure also concentrated on at most that many atoms due to the represented theorem. This makes the flow piecewise linear in time, and makes μ_t be concentrated on a finite number of atoms for all $t > 0$ due to the optimality condition. More work is needed to determine whether these effects combined can be used within this framework to provide a better discretization of the parameter space Ω or avoid the discretization all together in order to beat the curse of dimensionality.

In this work, we focussed on finding a sparse representation of a Barron space function. These functions represent shallow neural networks. For deep networks, there are several candidates spaces, like the Tree-like spaces [25], the Neural Hilbert Ladders [20], the Deep Variational Splines [35] and the Deep RKBS [3]. These candidates spaces have different structures than the Barron space. For example, in the Tree-like spaces the measure is defined over an infinite dimensional parameter space Ω instead of a finite dimensional set, and for the Deep RKBS a sequence of measures has to be optimised over simultaneously instead of just one measure μ_t. These function spaces pose specific challenges that need to be addressed, before the method can be applied successfully. We leave the investigation to which of these structures the methods in this paper can be best extended to future work.

8 Conclusion

This paper investigates an inverse problem for neural networks in the infinite width limit, which is to find a sparse representation of a Barron function that fits the data well. Sparse neural networks are known to improve generalization from training to test data, yet existing methods for identifying infinite width neural networks do not guarantee sparse solutions. We propose solving the inverse problem using the inverse scale space flow. The study systematically analyzes the convergence properties of this flow under ideal conditions and in the presence of measurement noise, sampling bias, and discretization. In the ideal setting, the objective decreases monotonically at a rate of $\mathcal{O}(1/t)$ to a minimizer, while in perturbed settings, convergence is achieved up to a bounded error. Discretized solutions are shown to converge to the full-space optimum when the mesh-size gets smaller.

Appendix A Notation and definitions

Let $\mathbb{R}$ denote the real numbers, and $\mathbb{N}$ denote the natural numbers without 0. The space of all Radon measures – regular, signed Borel measures with bounded total variation – on a locally compact Hausdorff Ω is denoted by $\mathcal{M}(\Omega)$. It is a Banach space with the norm

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

$$\|\mu\|_{\mathcal{M}(\Omega)} = \int_\Omega d|\mu|(x),$$

where $|\mu|$ is the total variation measure of μ. When Ω is compact and $\mathcal{M}(\Omega)$ is equipped with the weak*-topology, then $\mathcal{M}(\Omega)$ is dual to $\mathcal{C}(\Omega)$, the space of continuous functions on Ω. When Ω is unbounded, then it is dual to $\mathcal{C}_0(\Omega)$, the space of continuous functions on Ω that go to zero at infinity. All Radon measures $\mu \in \mathcal{M}(\Omega)$ have a polar decomposition, i.e. there exists a $\operatorname{sgn}\{\mu\} \in L^1(\Omega, |\mu|)$ with $|\operatorname{sgn}\{\mu\}| \leq 1$ such that

$$d\mu(x) = \operatorname{sgn}\{\mu\}(x) d|\mu|(x).$$

The space of all probability measures on a set U with finite k-th moments is denoted by $\mathcal{P}_k(U) \subseteq \mathcal{M}(U)$. The Wasserstein-1 metric between two probability measures $\rho, \pi \in \mathcal{P}_1(\Omega)$, can be computed by

$$W_1(\rho, \pi) = \sup\left\{ \int_\Omega f(\omega) d\rho(\omega) - \int_\Omega f(\omega) d\pi(\omega) \;\middle|\; f \in \mathcal{C}^0(\Omega),\ \operatorname{Lip}(f) \leq 1 \right\},$$

where $\operatorname{Lip}(f)$ denotes the Lipschitz constant of f. Given a set X, a positive number $p \in [1, \infty]$ and a radon measure $\rho \in \mathcal{M}(X)$, we write $L^p(\rho)$ instead of $L^p(X, \rho)$. If $f : L^\infty([0,\infty)) \to \mathcal{U}$ with $\mathcal{U}$ a normed vector space, then $f_t := f(t) \in \mathcal{U}$, f is Bochner integrable, and f has norm $\|f\|_{L^\infty([0,\infty),\mathcal{U})} = \operatorname{ess\,sup}_{t\in[0,\infty)} \|f_t\|_{\mathcal{U}}$. If $f : \mathcal{U} \to \mathcal{V}$ is an operator from $\mathcal{U}$ to $\mathcal{V}$, then the operator norm is denoted $\|f\|_{\mathcal{U}\to\mathcal{V}}$. If $U \subset V$ is a convex set, V is a locally convex space and $J : U \to \mathbb{R}$ is a convex function, then the convex conjugate is written as J^* and the subgradient ∂J of J at u_0 is given by

$$\partial J(u_0) = \{v \in V^* \mid J(u) - J(u_0) \geq \langle v \,|\, u - u_0 \rangle_{V^*},\ \forall\, u \in U\}.$$

(Fréchet) derivatives of a function or operator f are also denoted ∂f. If the derivative is a partial derivative, then a subscript will be added to indicate the variable with which the derivative is taken.

Acknowledgements

T. J. Heeringa and C. Brune acknowledge support by Sectorplan Bèta (the Netherlands) under the focus area "Mathematics of Computational Science". M. Burger, T. Roith and C. Brune acknowledge support of the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie (Grant No. 777826) (NoMADS). M. Burger and T. Roith further acknowledge support from DESY (Hamburg, Germany), a member of the Helmholtz Association HGF, by the German Ministry of Science and Technology (BMBF) (Grant No. 05M2020) (DELETO). M. Burger also acknowledges support from the German Research Foundation, Project BU 2327/19-1. Most of this study was carried out while T. Roith was affiliated with the Friedrich-Alexander-Universität Erlangen-Nürnberg.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS

References

[1] M. Bachmayr and M. Burger, Iterative total variation schemes for nonlinear inverse problems, *Inverse Problems*, 25(10):105004, 2009.

[2] F. Bartolucci, E. De Vito, L. Rosasco, and S. Vigogna, Understanding neural networks with reproducing kernel Banach spaces, *Appl. Comput. Harmon. Anal.*, 62:194–236, 2023.

[3] F. Bartolucci, E. De Vito, L. Rosasco, and S. Vigogna, Neural reproducing kernel Banach spaces and representer theorems for deep networks, *arXiv:2403.08750*, 2024.

[4] A. Beck and M. Teboulle, Mirror descent and nonlinear projected subgradient methods for convex optimization, *Oper. Res. Lett.*, 31(3):167–175, 2003.

[5] M. Benning, M. M. Betcke, M. J. Ehrhardt, and C.-B. Schönlieb, Choose your path wisely: Gradient descent in a Bregman distance framework, *SIAM J. Imaging Sci.*, 14(2):814–843, 2021.

[6] M. Benning and M. Burger, Modern regularization methods for inverse problems, *Acta Numer.*, 27:1–111, 2018.

[7] A. Braides, *A Handbook of* Γ*-Convergence*, in: *Handbook of Differential Equations. Stationary Partial Differential Equations*, Vol. 3, Elsevier, 2006.

[8] K. Bredies and H. K. Pikkarainen, Inverse problems in spaces of measures, *ESAIM Control Optim. Calc. Var.*, 19(1):190–218, 2013.

[9] H. Brézis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions Dans Les Espaces de Hilbert*, North-Holland Pub. Co., 1973.

[10] H. Brezis, Monotone operators, nonlinear semigroups and applications, in: *Proceedings of International Congress of Mathematicians*, 249–255, Canadian Math. Congress, 1974.

[11] C. Brune, A. Sawatzky, and M. Burger, Primal and dual Bregman methods with application to optical nanoscopy, *Int. J. Comput. Vision*, 92(2):211–229, 2011.

[12] L. Bungert, T. Roith, D. Tenbrinck, and M. Burger, Neural architecture search via Bregman iterations, *arXiv:2106.02479*, 2021.

[13] L. Bungert, T. Roith, D. Tenbrinck, and M. Burger, A Bregman learning framework for sparse neural networks, *J. Mach. Learn. Res.*, 23(1):8673–8715, 2022.

[14] M. Burger, G. Gilboa, S. Osher, and J. Xu, Nonlinear inverse scale space methods, *Commun. Math. Sci.*, 4(1):179–212, 2006.

[15] M. Burger, M. Möller, M. Benning, and S. Osher, An adaptive inverse scale space method for compressed sensing, *Math. Comp.*, 82(281):269–299, 2012.

[16] M. Burger and S. Osher, Convergence rates of convex variational regularization, *Inverse Problems*, 20(5):1411–1421, 2004.

[17] M. Burger, E. Resmerita, and L. He, Error estimation for Bregman iterations and inverse scale space methods in image restoration, *Computing*, 81(2-3):109–135, 2007.

[18] J.-F. Cai, S. Osher, and Z. Shen, Linearized Bregman iterations for compressed sensing, *Math. Comp.*, 78(267):1515–1536, 2009.

[19] J.-F. Cai, S. Osher, and Z. Shen, Convergence of the linearized Bregman iteration for l_1-norm minimization, *Math. Comp.*, 78(268):2127–2136, 2009.

[20] Z. Chen, Neural Hilbert ladders: Multi-layer neural networks in function space, *J. Mach. Learn. Res.*, 25(109):1–65, 2024.

[21] L. Chizat and F. Bach, On the global convergence of gradient descent for over-parameterized models using optimal transport, in: *Advances in Neural Information Processing Systems*, Vol. 31, Curran Associates, Inc., 2018.

[22] X. Dai, H. Yin, and N. K. Jha, NeST: A neural network synthesis tool based on a grow-and-prune paradigm, *IEEE Trans. Comput.*, 68(10):1487–1497, 2019.

[23] L. Devroye, L. Györfi, G. Lugosi, and H. Walk, On the measure of Voronoi cells, *J. Appl. Probab.*, 54(2):394–408, 2017.

[24] W. E, C. Ma, and L. Wu, The Barron space and the flow-induced function spaces for neural network models, *Constr. Approx.*, 55(1):369–406, 2022.

[25] W. E and S. Wojtowytsch, On the Banach spaces associated with multi-layer ReLU networks: Function

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48

OPEN ACCESS

representation, approximation theory and gradient descent dynamics, *arXiv:2007.15623*, 2020.
[26] W. E and S. Wojtowytsch, Representation formulas and pointwise properties for Barron functions, *Calc. Var. Partial Differ. Equations*, 61(2):46, 2022.
[27] T. J. Heeringa, L. Spek, F. L. Schwenninger, and C. Brune, Embeddings between Barron spaces with higher-order activation functions, *Appl. Comput. Harmon. Anal.*, 73:101691, 2024.
[28] S. Liu, D. C. Mocanu, A. R. R. Matavalam, Y. Pei, and M. Pechenizkiy, Sparse evolutionary deep learning with over one million artificial neurons on commodity hardware, *Neural Comput. Appl.*, 33(7):2589–2604, 2021.
[29] S. Liu, D. C. Mocanu, and M. Pechenizkiy, Intrinsically sparse long short-term memory networks, *arXiv:1901.09208*, 2019.
[30] M. Moeller and M. Burger, Multiscale methods for polyhedral regularizations, *SIAM J. Optim.*, 23(3): 1424–1456, 2013.
[31] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, Pruning convolutional neural networks for resource efficient inference, *arXiv:1611.06440*, 2017.
[32] Y. Nesterov, A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$, *Doklady AN SSSR*, 269(3):543–547, 1983.
[33] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, An iterative regularization method for total variation-based image restoration, *Multiscale Model. Simul.*, 4(2):460–489, 2005.
[34] R. Parhi and R. D. Nowak, Banach space representer theorems for neural networks and ridge splines, *J. Mach. Learn. Res.*, 22(1):1960–1999, 2021.
[35] R. Parhi and R. D. Nowak, What kinds of functions do deep neural networks learn? Insights from variational spline theory, *SIAM J. Math. Data Sci.*, 4(2):464–489, 2022.
[36] V. Ramanujan, M. Wortsman, A. Kembhavi, A. Farhadi, and M. Rastegari, What's hidden in a randomly weighted neural network?, in: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11890–11899, IEEE, 2020.
[37] J. Shi and S. Osher, A nonlinear inverse scale space method for a convex multiplicative noise model, *SIAM J. Imaging Sci.*, 1(3):294–321, 2008.
[38] L. Spek, T. J. Heeringa, F. Schwenninger, and C. Brune, Duality for neural networks through reproducing kernel Banach spaces, *arXiv:2211.05020*, 2023.
[39] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Stat. Soc., B: Stat. Methodol.*, 58(1):267–288, 1996.
[40] X. Wang and M. Benning, A lifted Bregman formulation for the inversion of deep neural networks, *Front. Appl. Math. Stat.*, 9:1176850, 2023.
[41] X. Wang and M. Benning, Lifted Bregman training of neural networks, *J. Mach. Learn. Res.*, 24(1):10909–10959, 2023.
[42] S. Wojtowytsch, On the convergence of gradient descent training for two-layer ReLU-networks in the mean field regime, *arXiv:2005.13530*, 2020.
[43] W. Yin, Analysis and generalizations of the linearized Bregman method, *SIAM J. Imaging Sci.*, 3(4):856–877, 2010.
[44] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing, *SIAM J. Imaging Sci.*, 1(1):143–168, 2008.

DOI https://doi.org/10.4208/jml.240123 | Generated on 2025-04-05 21:33:48
OPEN ACCESS