PREDICTIVE ONLINE OPTIMISATION WITH APPLICATIONS TO OPTICAL FLOW

Tuomo Valkonen*

Abstract Online optimisation revolves around new data being introduced into a problem while it is still being solved; think of deep learning as more training samples become available. We adapt the idea to dynamic inverse problems such as video processing with optical flow. We introduce a corresponding *predictive online primal-dual proximal splitting* method. The video frames now *exactly correspond to the algorithm iterations*. A user-prescribed predictor describes the evolution of the primal variable. To prove convergence we need a predictor for the dual variable based on (proximal) gradient flow. This affects the model that the method asymptotically minimises. We show that for inverse problems the effect is, essentially, to construct a new dynamic regulariser based on infimal convolution of the static regularisers with the temporal coupling. We finish by demonstrating excellent real-time performance of our method in computational image stabilisation and convergence in terms of regularisation theory.

1 INTRODUCTION

On Hilbert spaces X_k and Y_k, ($k \in \mathbb{N}$), consider the formal problem

$$\min_{x^1,x^2,\dots} \sum_{k=1}^{\infty} F_k(x^k) + G_k(K_k x^k) \quad \text{s.t.} \quad x^{k+1} = \bar{A}_k(x^k), \tag{1.1}$$

where $F_k : X_k \to \overline{\mathbb{R}}$ and $G_k : Y_k \to \overline{\mathbb{R}}$ are convex, proper, and lower semicontinuous, $K_k \in \mathbb{L}(X_k; Y_k)$ is linear and bounded, and the *temporal coupling* operators $\bar{A}_k : X_k \to X_{k+1}$. One may think of $\min(F_k + G_k \circ K_k)$ as a problem we want to solve on each time instant k, knowing that the solutions of these problems are coupled via the environment acting through $\bar{A}_k$. For example, $\bar{A}_k$ can describe the true movement of objects in a scene, that we cannot control, and do not necessarily know. This problem is clearly challenging; even its solutions are generally well-defined only asymptotically.

Instead of trying to solve (1.1) exactly, what if we take only *one step* of an optimisation algorithm on each partial problem

$$\min_{x^k \in X_k} J_k(x_k) := F_k(x^k) + G_k(K_k x^k), \tag{1.2}$$

and use an approximation $A_k : X_k \to X_{k+1}$, called the *predictor*, of the unknown $\bar{A}_k$ to transfer iterates between the steps? Can we obtain convergence in an asymptotic sense, and to what? We set out to study these questions, in particular to develop a predictive "online" primal-dual method.

This research has been supported by Escuela Politécnica Nacional internal grant PIJ-18-03 and Academy of Finland grants 314701 and 320022.

*Department of Mathematics and Statistics, University of Helsinki, Finland *and* ModeMat, Escuela Politécnica Nacional, Quito, Ecuador, tuomo.valkonen@iki.fi, ORCID: 0000-0001-6683-3572

Our simple model problem is image sequence denoising: we are given noisy images $\{b^k\}_{k\in\mathbb{N}}$ in the space[1] $X = L^2(\Omega)$ on the two-dimensional domain $\Omega \subset \mathbb{R}^2$, and bijective displacement fields $v^k : \Omega \to \Omega$ such that the images roughly satisfy the *optical flow* constraint $b^{k+1} \approx A_k(b^k)$ for $A_k(x) := x \circ v^k$. For an introduction to optical flow, we refer to [5]. The static problem (1.2) is the isotropic total variation denoising

$$\min_{x\in X} \frac{1}{2}\|x - b^k\|_X^2 + \alpha\|Dx\|_{\mathcal{M}}, \tag{1.3}$$

where $\alpha > 0$ is a regularisation parameter and D a measure-valued differential operator. In the dynamic case we would like the approximate solutions $\{x^k\}_{k\in\mathbb{N}}$ to also satisfy $x^{k+1} \approx A_k(x^k)$. In principle, we could for the first N frames for some penalisation parameter $\beta > 0$ solve

$$\min_{x^1,\dots,x^{N+1}\in X} \sum_{k=0}^{N} \left(\frac{1}{2}\|x^k - b^k\|_X^2 + \alpha\|Dx^k\|_{\mathcal{M}} + \frac{\beta}{2}\|x^{k+1} - A_k(x^k)\|_X^2 \right),$$

or a version that linearises A_k. However, when the number of frames N is high, these problems become numerically increasingly challenging. Also, if we want to solve the problem for $N + 1$ frames, we may need to do the same amount of work again, depending on how well our algorithm can "restart". Primal-dual methods in particular tend to be very sensitive to initialisation.

An alternative is to try to solve the problem in an "online" fashion, building the gradually changing data into the algorithm design [38]. We refer to [19, 6, 25] for introductions and further references to online methods in machine learning. Online Newton methods have also been studied for smooth PDE-constrained optimisation [8, 17]. Our approach has more in common with machine learning and nonsmooth optimisation. From this point of view, basic online methods seek a low *regret* for a dynamic solution sequence compared to a fixed solution. With the notation $x^{1:N} := (x^1, \dots, x^N)$, for any comparison set $B \subset X$, where we expect the true solution to lie, we define the regret as

$$\operatorname{regret}_B(x^{1:N}) := \sup_{\bar{x}\in B} \sum_{k=1}^{N} \left(J_k(x^k) - J_k(\bar{x}) \right).$$

This does not model the temporal nature of our problem, so in [18] *dynamic regret* is introduced. For a *comparison set* $\mathcal{B}_{1:N} \subset \prod_{k=1}^N X_k$ of potential true solutions, it reads

$$\operatorname{dynamic\_regret}_{\mathcal{B}_{1:N}}(x^{1:N}) := \sup_{\bar{x}^{1:N}\in\mathcal{B}_{1:N}} \sum_{k=1}^{N} \left(J_k(x^k) - J_k(\bar{x}^k) \right). \tag{1.4}$$

For example, we can take

$$\mathcal{B}_{1:N} = \{(\bar{x}^1, \dots, \bar{x}^N) \mid \bar{x}^0 \in \mathcal{B}_0,\ \bar{x}^{k+1} = \bar{A}_k(\bar{x}^k),\ k = 0, \dots, N-1\} \tag{1.5}$$

for some $\mathcal{B}_0 \subset X_0$, where we expect the initial true $\bar{x}^0$ to lie, and the true temporal coupling operators $\bar{A}_k : X_k \to X_{k+1}$. For the optical flow problem, (1.5) would read

$$\mathcal{B}_{1:N} = \{(\bar{x}^0 \circ \bar{v}_1, \dots, \bar{x}^0 \circ \bar{v}_1 \circ \cdots \circ \bar{v}_N) \mid \bar{x}^0 \in \mathcal{B}_0\}$$

[1] The total variation term in (1.3) in principle requires $x \in \mathrm{BV}(\Omega)$, the space of functions of bounded variation on Ω. This is not a Hilbert space, but merely a Banach space, where our overall setup (1.1) does not to apply. However, due to the weak(-∗) lower semicontinuity of convex functionals, any minimiser of (1.3) necessarily lies in $L^2(\Omega) \cap \mathrm{BV}(\Omega)$, so we are justified in working in the Hilbert space $X = L^2(\Omega)$, and seeing $\mathrm{BV}(\Omega)$ as a constraint imposed by the total variation term.

for some true displacement fields $\bar{v}_k$ and a set $\mathcal{B}_0$ containing the initial non-corrupted frame $\bar{x}^0$. Thus $\mathcal{B}_{1:N}$ consists of all potential "true" frames $\bar{x}^1, \dots, \bar{x}^N$ generated from all potential initial $\bar{x}^0$ by the true displacement fields. When the dynamic regret (1.4) is below zero, the algorithmic iterates $x^{1:N}$ fit the data and total variation regularisation of (1.3) better than all $\bar{x}^{1:N} \in \mathcal{B}_{1:N}$, but may not satisfy the constraint $x^{1:N} = (x^0 \circ v_1, \dots, x^0 \circ v_1 \circ \cdots \circ v_N)$ for *any* displacement fields v_k. Specific algorithms may additionally seek to approximately satisfy this constraint for some measured or estimated displacement fields v_k.

The idea now would be to obtain a low dynamic regret by some strategy. One possibility is what we already mentioned: take one step of an optimisation method towards a minimiser of each J_k, and then use A_k to predict an approximate solution for the next problem. Repeat. In this approach, data frames exactly correspond to algorithm iterations. The strategy of very inexact solutions is motivated by the fact that neural networks can be effective—not get stuck in local optima—because subproblems are not solved exactly [9]. A different type of applications with only intermittent sampling is studied in [2, 27]

In Section 2 we we prove low dynamic regret for predictive forward-backward splitting, in line with the literature [18, 36]. This serves to introduce concepts and ideas for our main interest: primal-dual methods. Indeed, forward-backward splitting is poorly applicable to (1.3): the proximal step is just as expensive as the original problem. It is more effective on the dual problem, however, we are given a primal predictor A_k. Moreover, purely dual formulations are not feasible for deblurring and more complex inverse problems. A solution is to work with primal-dual formulations of the static problems (1.2),

$$\min_{x \in X_k} \max_{y \in Y_k} F_k(x) + \langle K_k x, y \rangle - G_k^*(y). \tag{1.6}$$

Here G_k^* is the Fenchel conjugate of G_k. A popular method for this type of problems is the *primal-dual proximal splitting* (PDPS) of Chambolle and Pock [11]. We refer to [31] for an overview of variants, alternatives, and extensions to non-convex problems.

MAIN CONTRIBUTIONS

We develop in Section 4 a predictive online PDPS for (1.1). For the primal variable we use the user-prescribed predictor $A_k : X_k \to X_{k+1}$, but for the dual variable the regret theory imposes a more technical predictor. This forms the main challenge of our work. To prepare for this, we introduce in Section 3 appropriate *partial primal gap* functionals to replace the dynamic regret (1.4), not applicable to primal-dual methods.

We finish in Section 5 with computational image stabilisation based on optical flow and online optimisation. We obtain real-time performance and show convergence of the algorithmic solutions in terms of regularisation theory [15] as the noise level decreases. Before this we introduce notation.

NOTATION

We write $x^{n:m} := (x^n, \dots, x^m)$ with $n \le m$, and $x^{n:\infty} := (x^n, x^{n+1}, \dots)$. We slice a set $\mathcal{B} \subset \prod_{k=0}^{\infty} X_k$ as $\mathcal{B}_{n:m} := \{x^{n:m} \mid x^{0:\infty} \in \mathcal{B}\}$ and $\mathcal{B}_n := \mathcal{B}_{n:n}$. We write $\mathbb{L}(X;Y)$ for the set of bounded linear operators between (Hilbert) spaces X and Y, and $\mathrm{Id} \in \mathbb{L}(X;X)$ for the identity operator. We write $\langle x, y \rangle_M := \langle Mx, y \rangle$ for $M \in \mathbb{L}(X;X)$ and, if M is positive semi-definite, also $\|x\|_M := \sqrt{\langle x, x \rangle_M}$.

We write $M \ge 0$ if M is positive semidefinite and $M \simeq N$ if $\langle Mx, x \rangle = \langle Nx, x \rangle$ for all x.

For any $A \subset X$ and $x \in X$ we set $\langle A, x \rangle := \{\langle z, x \rangle \mid z \in A\}$. We write δ_A for the $\{0, \infty\}$-valued indicator function of A. For any $B \subset \mathbb{R}$ (in particular $B = \langle A, x \rangle$), we use the notation $B \ge 0$ to mean that $t \ge 0$ for all $t \in B$.

For $F : X \to (-\infty, \infty]$, we write $\operatorname{dom} F := \{x \in X \mid F(x) < \infty\}$ for the effective domain. With $\overline{\mathbb{R}} := [\infty, \infty]$ the set of extended reals, we call $F : X \to \overline{\mathbb{R}}$ *proper* if $F > -\infty$ and $\operatorname{dom} F \neq \emptyset$. Let

then F be convex. We write $\partial F(x)$ for the subdifferential at x and (for additionally proper and lower semicontinuous F)

$$\operatorname{prox}_F(x) := \operatorname*{arg\,min}_{\tilde{x}\in X} \; F(\tilde{x}) + \frac{1}{2}\|\tilde{x} - x\|^2 = (\mathrm{Id} + \partial F)^{-1}(x)$$

for the proximal map. We recall that F is strongly subdifferentiable at x with the factor $\gamma > 0$ if

$$F(\tilde{x}) - F(x) \geq \langle z, \tilde{x} - x\rangle + \frac{\gamma}{2}\|\tilde{x} - x\|^2 \quad \text{for all} \quad z \in \partial F(x) \text{ and } \tilde{x} \in X.$$

In Hilbert spaces this is equivalent to strong convexity with the same factor.

Finally, for $f \in L^q(\Omega; \mathbb{R}^n)$, we write $\|f\|_{p,q} := \big\|\xi \mapsto \|f(\xi)\|_p\big\|_{L^q(\Omega)}$.

2 PREDICTIVE ONLINE FORWARD-BACKWARD SPLITTING

We review predictive online forward-backward splitting (POFB) for (1.1) with $K_k = \mathrm{Id}$. This is useful to explain online methods in general and to motivate our proofs and the dual comparison sequence for the online PDPS. We recall that given a step length parameter $\tau > 0$, forward-backward splitting for $\min[F + G]$ iterates

$$x^{k+1} := \operatorname{prox}_{\tau G}(x^k - \tau \nabla F(x^k)).$$

We present a predictive online version in Algorithm 2.1. To study it, we work with:

Assumption 2.1. For all $k \geq 1$: $F_k, G_k : X_k \to \mathbb{R}$ are convex, proper, and lower semicontinuous on a Hilbert space X_k. ∇F_k exists and is L_k-Lipschitz. We write $J_k := F_k + G_k$ and $\gamma_{F_k}, \gamma_{G_k} \geq 0$ for the factors of (strong) subdifferentiability of F_k and G_k. We suppose for some step length parameters $\tau_k > 0$ and some $\zeta_k \in (0, 1]$ that

$$0 \leq \gamma_k := \begin{cases} \gamma_{G_k} + \gamma_{F_k} - \tau_k \zeta_k^{-1} L_k^2, & \gamma_{F_k} > 0, \\ \gamma_{G_k}, & \gamma_{F_k} = 0 \text{ in which case we require } \tau_k L_k \leq \zeta_k. \end{cases} \tag{2.1}$$

We are also given predictors $A_k : X_k \to X_{k+1}$ and a bounded comparison set $\mathcal{B} \subset \prod_{k=0}^{\infty} X_k$ of potential true solutions. They satisfy for some (Lipschitz-like) factor Λ_k and *prediction error* ε_{k+1} the *prediction bound*

$$\frac{1}{2}\|A_k(x^k) - \bar{x}^{k+1}\|^2 \leq \frac{\Lambda_k}{2}\|x^k - \bar{x}^k\|^2 + \varepsilon_{k+1} \quad (\bar{x}^{0:\infty} \in \mathcal{B},\ k \in \mathbb{N}). \tag{2.2}$$

Remark 2.2. Typically $\mathcal{B}$ is given as in (1.5) by some true (unknown) temporal coupling operators $\bar{A}_k : X_k \to X_{k+1}$ that the (known) predictors A_k approximate. Then (2.2) reads

$$\frac{1}{2}\|A_k(x^k) - \bar{A}_k(\bar{x}^k)\|^2 \leq \frac{\Lambda_k}{2}\|x^k - \bar{x}^k\|^2 + \varepsilon_{k+1}.$$

If we knew that $\bar{A}_k = A_k$, and the operator were Lipschitz, we could take Λ_k as the Lipschitz factor and the prediction error $\varepsilon_{k+1} = 0$. Typically, however, we would not know the true temporal coupling—or would know it only up to measurement noise—so need the prediction errors to model this lack of knowledge or noise.

We need to develop regret theory for Algorithm 2.1. We recall the following *smoothness three-point inequalities* found in, e.g., [30, Appendix B] and [14, Chapter 7].

Algorithm 2.1 Predictive online forward-backward splitting (POFB)

Require: For all $k \in \mathbb{N}$, on Hilbert spaces X_k, a primal predictor $A_k : X_k \to X_{k+1}$ and convex, proper, lower semicontinuous $F_{k+1}, G_{k+1} : X_{k+1} \to \overline{\mathbb{R}}$ such that G_{k+1} has Lipschitz gradient. Step length parameters $\tau_{k+1} > 0$.

1: Pick an initial iterate $x^0 \in X_0$.
2: **for** $k \in \mathbb{N}$ **do**
3: $\quad z^{k+1} := A_k(x^k)$ ▷ prediction
4: $\quad x^{k+1} := \operatorname{prox}_{\tau_{k+1} G_{k+1}}(z^{k+1} - \tau_{k+1} \nabla F_{k+1}(z^{k+1}))$ ▷ forward-backward step
5: **end for**

Lemma 2.3. *Suppose $F : X \to \overline{\mathbb{R}}$ is convex, proper, and lower semicontinuous, and has L-Lipschitz gradient. Then*

$$\langle \nabla F(z), x - \bar{x} \rangle \geq F(x) - F(\bar{x}) - \frac{L}{2}\|x - z\|^2 \quad (\bar{x}, z, x \in X). \tag{2.3}$$

If F is, moreover, γ-strongly convex, then for any $\beta > 0$,

$$\langle \nabla F(z), x - \bar{x} \rangle \geq F(x) - F(\bar{x}) + \frac{\gamma - \beta L^2}{2}\|x - \bar{x}\|^2 - \frac{1}{2\beta}\|x - z\|^2 \quad (\bar{x}, z, x \in X). \tag{2.4}$$

Lemma 2.4. *Suppose Assumption 2.1 holds. Then, for any $k \in \mathbb{N}$,*

$$\langle \partial G_k(x^k) + \nabla F_k(z^k), x^k - \bar{x}^k \rangle \geq J_k(x^k) - J_k(\bar{x}^k) + \frac{\gamma_k}{2}\|x^k - \bar{x}^k\|^2 - \frac{\zeta_k}{2\tau_k}\|x^k - z^k\|^2.$$

Proof. If $\gamma_{F_k} = 0$, (2.3) in Lemma 2.3 with the (strong) subdifferentiability of G_k yield

$$\langle \partial G_k(x^k) + \nabla F_k(z^k), x^k - \bar{x}^k \rangle \geq J_k(x^k) - J_k(\bar{x}^k) + \frac{\gamma_{G_k}}{2}\|x^k - \bar{x}^k\|^2 - \frac{L_k}{2}\|x^k - z^k\|^2.$$

Due to (2.1) and Assumption 2.1 ensuring $\tau_k L_k \leq \zeta_k$, this gives the claim in the case $\gamma_{F_k} = 0$.

If $\gamma_{F_k} > 0$, by (2.4) for $\beta = \zeta_k^{-1}\tau_k$ and the (strong) subdifferentiability of G_k,

$$\begin{aligned}\langle \partial G_k(x^k) + \nabla F_k(z^k), x^k - \bar{x}^k \rangle \geq{}& J_k(x^k) - J_k(\bar{x}^k) \\ &+ \frac{\gamma_{G_k} + \gamma_{F_k} - \zeta_k^{-1}\tau_k L_k^2}{2}\|x^k - \bar{x}^k\|^2 - \frac{\zeta_k}{2\tau_k}\|x^k - z^k\|^2.\end{aligned}$$

This gives the claim by the case $\gamma_{F_k} > 0$ of (2.1). □

We now have the tools to study regret. The sets $\mathcal{B}_{1:N}$ in the following results would typically be given by (1.5) through some true temporal coupling operators $\bar{A}_k : X_k \to X_{k+1}$. The "testing parameters" φ_k can be used to derive regret rates from the regularity of the problem. We explain them in the corollary and remark to follow.

Theorem 2.5. *Suppose Assumption 2.1 holds and some* testing parameters $\{\varphi_k\}_{k\in\mathbb{N}} \subset \mathbb{R}$ *satisfy $\varphi_{k+1} \leq \varphi_k(1 + \gamma_k\tau_k)\Lambda_k^{-1}$ for all $k = 0, \ldots, N-1$. Let $x^{1:N}$ generated by Algorithm 2.1 for an $x^0 \in X_0$. Then*

$$\begin{aligned}\sup_{\bar{x}^{1:N} \in \mathcal{B}_{1:N}} \sum_{k=1}^{N} \varphi_k \tau_k [J_k(x^k) - J_k(\bar{x}^k)] + \sum_{k=0}^{N-1} \frac{\varphi_{k+1}(1 - \zeta_{k+1})}{2}\|x^{k+1} - A_k(x^k)\|^2 \\ \leq \sup_{\bar{x}^0 \in \mathcal{B}_0} \frac{\varphi_0(1 + \gamma_0\tau_0)}{2}\|x^0 - \bar{x}^0\|^2 + \sum_{k=1}^{N} \varepsilon_k \varphi_k.\end{aligned} \tag{2.5}$$

Proof. We can write Algorithm 2.1 implicitly as

$$0 \in \tau_k[\partial G_k(x^k) + \nabla F_k(z^k)] + (x^k - z^k) \quad (k = 1, \dots, N) \tag{2.6}$$

where $z^{k+1} := A_k(x^k)$ for $k = 0, \dots, N-1$. Following the testing methodology of [30, 14], we take any $\bar{x}^k \in X_k$ and apply the linear "testing operator" $\varphi_k\langle \cdot, x^k - \bar{x}^k\rangle$ to both sides of (2.6). Following with Lemma 2.4, this yields

$$0 \geq \varphi_k\langle x^k - z^k, x^k - \bar{x}^k\rangle + \frac{\varphi_k\gamma_k\tau_k}{2}\|x^k - \bar{x}^k\|^2 - \frac{\varphi_k}{2}\|x^k - z^k\|^2 + \mathcal{G}_k \quad (k = 1, \dots, N) \tag{2.7}$$

for

$$\mathcal{G}_k := \varphi_k\tau_k[J_k(x^k) - J_k(\bar{x}^k)] + \frac{\varphi_k(1-\zeta_k)}{2}\|x^k - z^k\|^2.$$

We recall the Pythagoras' identity or three-point formula

$$\langle x^k - z^k, x^k - \bar{x}^k\rangle = \frac{1}{2}\|x^k - z^k\|^2 - \frac{1}{2}\|z^k - \bar{x}^k\|^2 + \frac{1}{2}\|x^k - \bar{x}^k\|^2. \tag{2.8}$$

Hence (2.7) yields

$$\frac{\varphi_k}{2}\|z^k - \bar{x}^k\|^2 \geq \frac{\varphi_k(1+\tau_k\gamma_k)}{2}\|x^k - \bar{x}^k\|^2 + \mathcal{G}_k \quad (k = 1, \dots, N).$$

Now taking $\bar{x}^{1:N} \in \mathcal{B}_{1:N}$ and using the prediction bound (2.2) followed by $\varphi_{k+1}\Lambda_k \leq \varphi_k(1+\gamma_k\tau_k)$, we obtain

$$\frac{\varphi_k(1+\gamma_k\tau_k)}{2}\|x^k - \bar{x}^k\|^2 + \varphi_{k+1}\varepsilon_{k+1} \geq \frac{\varphi_{k+1}(1+\gamma_{k+1}\tau_{k+1})}{2}\|x^{k+1} - \bar{x}^{k+1}\|^2 + \mathcal{G}_{k+1} \quad (k = 0, \dots, N-1).$$

Now we just sum over $k = 0, \dots, N-1$ and take the supremum over $\bar{x}^{1:N} \in \mathcal{B}_{1:N}$. □

The next corollary, obtained with $\varphi_k \equiv 1$ and constant $\tau_k \equiv \tau$, is similar to [18, Theorem 4] in the case $1 + \gamma_k\tau \geq \Lambda_k$, i.e., when any available strong convexity balances the non-expansivity-like $\Lambda_k > 1$ in the prediction bound (2.2). Often in the online optimisation literature, $\mathrm{regret}_B(x^1, \dots, x^N) \leq C\sqrt{N}$. The growing regret bound can arise from violating this step length condition or from the penalties $\sum_{k=1}^N \varepsilon_k$ in the prediction bound (2.2). For our purposes, bounding the regret in terms of the initialisation and the prediction bounds is enough.

Corollary 2.6. *Suppose Assumption 2.1 holds with $\tau_k \equiv \tau$ and $1 + \gamma_k\tau \geq \Lambda_k$ for all $k = 0, \dots, N-1$. Let $x^{1:N}$ generated by Algorithm 2.1 for an initial $x^0 \in X$. Then*

$$\mathrm{dynamic\_regret}_{\mathcal{B}_{1:N}}(x^1, \dots, x^N) + \sum_{k=0}^{N-1}\frac{1-\zeta_{k+1}}{2\tau}\|x^{k+1} - A_k(x^k)\|^2 \leq \sup_{\bar{x}^0 \in \mathcal{B}_0}\frac{\|x^0 - \bar{x}^0\|^2}{2\tau(1+\gamma_0\tau)^{-1}} + \sum_{k=1}^N \frac{\varepsilon_k}{\tau}.$$

Remark 2.7 (Weighted dynamic regret). Suppose $1 + \gamma_k\tau_k > \Lambda_k$. Then $\{\varphi_k\}_{k\in\mathbb{N}}$ can increase while satisfying $\varphi_{k+1} \leq \varphi_k(1+\gamma_k\tau_k)\Lambda_k^{-1}$. If $\inf_k \tau_k > 0$, then (2.5) places more importance on J_k for large k: we regret early iterates less than recent. If $\frac{1+\gamma_k\tau_k}{\Lambda_k} \geq c > 1$ and $\varphi_k = c^k\varphi_0$, this growth in importance is exponential, comparable to linear convergence on static problems; cf. [30]. With $F_{k+1} \equiv 0$ it is even possible to take $\tau_k \nearrow \infty$ and obtain superexponential growth (superlinear convergence).

If, on the other hand $1 + \gamma_k\tau_k < \Lambda_k$, then the condition $\varphi_{k+1} \leq \varphi_k(1+\gamma_k\tau_k)\Lambda_k^{-1}$ forces $\{\varphi_k\}_{k\in\mathbb{N}}$ to be decreasing. We therefore regret bad early iterates more than the recent. In the context of static optimisation problems, we are in the region of non-convergence or at most slow sub-$O(1/N)$ rates.

3 PARTIAL GAP FUNCTIONALS

We start our development of a primal-dual method by deriving meaningful measures of regret. We cannot in general obtain estimates on conventional duality gaps or on iterates, so need alternative criteria. Throughout this section $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ are convex, proper, and lower semicontinuous, and $K \in \mathbb{L}(X; Y)$ on Hilbert spaces X and Y. We write $\mathcal{L}(x, y) := F(x) + \langle Kx, y\rangle - G^*(y)$ for the corresponding *Lagrangian*. We recall that the first-order primal-dual optimality conditions for

$$\min_{x\in X} F(x) + G(Kx) \quad \text{equiv.} \quad \min_{x\in X}\max_{y\in Y} \mathcal{L}(x, y)$$

are

$$-K\hat{y} \in \partial F(\hat{x}) \quad \text{and} \quad K^*\hat{x} \in \partial G^*(\hat{y}). \tag{3.1}$$

We call such a pair $(\hat{x}, \hat{y})$ a *critical point*.

3.1 COMMON GAP FUNCTIONALS

By the Fenchel–Young inequality applied to $T(x, y) := F(x) + G^*(y)$, the *duality gap*

$$\mathcal{G}(x, y) := [F(x) + G(Kx)] + [F^*(-K^*y) + G^*(y)] \geq 0,$$

and is zero if and only if (3.1) holds. We can expand

$$\mathcal{G}(x, y) = \sup_{(\bar{x},\bar{y})\in X\times Y} \left(\mathcal{L}(x, \bar{y}) - \mathcal{L}(\bar{x}, y)\right).$$

This motivates the *Lagrangian duality gap*

$$\mathcal{G}^{\mathcal{L}}(x, y; \bar{x}, \bar{y}) := \mathcal{L}(x, \bar{y}) - \mathcal{L}(\bar{x}, y).$$

It is non-negative if $(\bar{x}, \bar{y})$ is a critical point, but may be zero even if (x, y) is not.

Since the Lagrangian duality gap is a relatively weak measure of optimality, and the true duality gap may not converge (fast), we define for bounded $B \subset X \times Y$ the *partial duality gap*

$$\mathcal{G}_B(x, y) := \sup_{(\bar{x},\bar{y})\in B} [\mathcal{L}(x, \bar{y}) - \mathcal{L}(\bar{x}, y)].$$

This is non-negative if B contains a critical point and equals the true duality gap $\mathcal{G}$ if $B = X \times Y$. The partial gap converges ergodically for the basic unaccelerated PDPS [11].

3.2 PARTIAL PRIMAL GAPS

If we are not interested in the dual variable, we can define the *partial primal gap*

$$\hat{\mathcal{G}}_B(x) := \sup_{(\bar{x},\bar{y})\in B} \inf_{y\in Y} \mathcal{G}^{\mathcal{L}}(x, y; \bar{x}, \bar{y}). \tag{3.2}$$

We now try to interpret it.

Lemma 3.1. *Let $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous, and $K \in \mathbb{L}(X; Y)$. Pick $B \subset X \times Y$. Then*

$$\hat{\mathcal{G}}_B(x) = [F + \check{G} \circ K](x) - \inf_{(\bar{x},\bar{y})\in B} [F + G \circ K](\bar{x}) \tag{3.3}$$

for

$$\check{G}(y') := \sup_{\bar{x}\in X,\bar{y}\in Y} \left(\langle y', \bar{y}\rangle - G^*(\bar{y}) - J_B(\bar{x}, \bar{y})\right) - J_B^*(0, 0) \quad \textit{and} \tag{3.4}$$

$$J_B(\tilde{x}, \tilde{y}) := F(\tilde{x}) + G(K\tilde{x}) + \delta_B(\tilde{x}, \tilde{y}).$$

Proof. We have

$$\begin{aligned}\inf_{y\in Y} \mathcal{G}^{\mathcal{L}}(x, y; \bar{x}, \bar{y}) &= \mathcal{L}(x, \bar{y}) - \sup_{y\in Y} \mathcal{L}(\bar{x}, y) \\ &= F(x) + \langle Kx, \bar{y}\rangle - G^*(\bar{y}) - [F + G \circ K](\bar{x}).\end{aligned}$$

Thus

$$\begin{aligned}\hat{\mathcal{G}}_B(x) &= F(x) + \sup_{\bar{x}\in X, \bar{y}\in Y} \left(\langle Kx, \bar{y}\rangle - G^*(\bar{y}) - [F + G \circ K](\bar{x}) - \delta_B(\bar{x}, \bar{y})\right) \\ &= F(x) + \sup_{\bar{x}\in X, \bar{y}\in Y} \left(\langle Kx, \bar{y}\rangle - G^*(\bar{y}) - J_B(\bar{x}, \bar{y})\right) = F(x) + \breve{G}(Kx) + J_B^*(0, 0).\end{aligned}$$

Since $J_B^*(0, 0) = -\inf_{(\bar{x},\bar{y})\in B}[F + G \circ K](\bar{x})$, this establishes the claim. □

Example 3.2. If $B = B_X \times Y$ for some $B_X \subset X$, then $J_B(x, y)$ does not depend on y so that we obtain $\breve{G} = G$. Thus the partial primal gap reduces to a standard difference of function values,

$$\hat{\mathcal{G}}_{B_X\times Y}(x) = [F + G \circ K](x) - \inf_{\bar{x}\in B_X} [F + G \circ K](\bar{x}).$$

If now B_X contains a minimiser of $F + G \circ K$, this difference is non-negative.

This example gives an indication towards the meaningfulness of the partial primal gap. In particular, if we take a smaller set B than in the example, we can expect $\hat{\mathcal{G}}_B(x)$ to attain smaller values. It may be negative even if B_X contains a minimiser of $F + G \circ K$. This is akin to the regret functionals from the Introduction. Indeed, we will use the partial primal gap as the basis for a *marginalised primal regret* that "fails to regret" what $F + \breve{G} \circ K \leq F + G \circ K$ cannot measure.

In the applications of Section 5, $G(y^1, \dots, y^N) = \sum_{k=1}^N \alpha\|Dy^k\|_{\mathcal{M}}$, compare (1.3), and B is a primal-dual extension $\mathcal{U}_{1:N}$ of $\mathcal{B}_{1:N}$ from (1.5). The construction of $\breve{G}$ convolves the static total variation regulariser G with the temporally coupled objective $J_{\mathcal{U}_{1:N}}$. The effect is to produce a new dynamic regulariser, alternative to [21, 35, 24, 23, 12, 26, 34]. The following instructive proposition elucidates how this works in general. However, the convexity assumption on B is not satisfied by $\mathcal{U}_{1:N}$. We write $E \,\square\, \tilde{E}$ for the infimal convolution of $E, \tilde{E} : X \to \overline{\mathbb{R}}$.

Proposition 3.3. *Suppose B is closed, convex, and nonempty, and both G and J_B are coercive. Then*

$$\breve{G}(y') = \inf_{\tilde{y}\in Y} \left(G(y' - \tilde{y}) + J_B^*(0, \tilde{y}) - J_B^*(0, 0)\right).$$

Proof. We recall that $(E \,\square\, \tilde{E})^* = E^* + \tilde{E}^*$ for proper $E, \tilde{E} : X \to \overline{\mathbb{R}}$ [3, Proposition 13.21]. The infimal convolution $E \,\square\, \tilde{E}$ is convex, proper, and lower semicontinuous when E and $\tilde{E}$ also are, E is coercive, and $\tilde{E}$ is bounded from below [3, Propositions 12.14]. Since then $(E \,\square\, \tilde{E})^{**} = E \,\square\, \tilde{E}$, we obtain $E \,\square\, \tilde{E} = (E^* + \tilde{E}^*)^*$.

By the convexity of B, $J_B = J_B^{**}$. The coercivity of J_B implies that J_B^* is bounded from below.[2] Since G is coercive, taking $E(x, y) = G(y) + \delta_{\{0\}}(x)$ and $\tilde{E} = J_B^*$, we get

$$\begin{aligned}\breve{G}(y') &= \sup_{\tilde{x}\in X, \tilde{y}\in Y} \left(\langle y', \tilde{y}\rangle - G^*(\tilde{y}) - J_B^{**}(\tilde{x}, \tilde{y})\right) - J_B^*(0, 0) \\ &= ([(\tilde{x}, \tilde{y}) \mapsto G^*(\tilde{y})]^* \,\square\, J_B^*)(0, y') - J_B^*(0, 0) \\ &= ([(\tilde{x}, \tilde{y}) \mapsto G(\tilde{y}) + \delta_{\{0\}}(\tilde{x})] \,\square\, J_B^*)(0, y') - J_B^*(0, 0) \\ &= \inf_{\tilde{y}\in Y} \left(G(y' - \tilde{y}) + J_B^*(0, \tilde{y}) - J_B^*(0, 0)\right). \quad \square\end{aligned}$$

[2] Any coercive, convex, proper, lower semicontinuous function $E : X \to \overline{\mathbb{R}}$ has a minimiser $\hat{x}$. By the Fermat principle $0 \in \partial E(\hat{x})$. Thus $\hat{x} \in \partial E^*(0)$, which says exactly that $E^* \geq E^*(0)$.

Example 3.4. Take $B = B_X \times B_Y$ for some convex and closed $B_X \subset X$ and $B_Y \subset Y$. Then Proposition 3.3 gives $\breve{G}(y') = (G \,\square\, \delta^*_{B_Y})(y')$.

In particular, let $G = \alpha\|\cdot\|_Y$ for some $\alpha > 0$ and $B_Y = B(\hat{y}, \rho) := \{y \in Y \mid \|y - \hat{y}\|_Y \leq \rho\}$ for some "expected solution" $\hat{y}$ and "confidence" $\rho > 0$. Then $\breve{G}(y') = (G^* + \delta_{B_Y})^*(y') = \delta^*_{B(0,\alpha)\cap B(\hat{y},\rho)}(y')$. If $\|\hat{y}\|_Y = \alpha$ and $\rho < \alpha$, this means that $\breve{G}$ will not penalise points $y' = Kx$ with $\langle y', \hat{y}\rangle \leq 0$. We might interpret this as follows: since we are highly confident (small ρ) that $Kx \propto \hat{y}$ for an optimal x, we are not even interested in studying dual variables that point in the opposite direction. If K were additionally a (discretised) gradient operator, as for total variation regularisation, roughly speaking this would say that we are not interested in studying gradients that point away from the expected gradient.

More generally, we can construct an infimal convolution lower bound with respect to the set of primal-dual minimisers of J_B. The coercivity assumption in the next lemma is fulfilled for F the squared distance or B bounded, both of which will be the case for the optical flow example.

Proposition 3.5. *Let $F : X \to \overline{\mathbb{R}}$ and $G : Y \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous, and $K \in \mathbb{L}(X;Y)$. Pick a closed subset $B \subset X \times Y$ and suppose J_B constructed from these components is coercive. Let*

$$\hat{B} := \{(\bar{x}, \bar{y}) \in B \mid J_B(\bar{x}, \bar{y}) = \inf J_B\} \quad \text{and} \quad \hat{B}_Y := \{\hat{y} \mid (\hat{x}, \hat{y}) \in \hat{B}\}.$$

Then $\breve{G}$ defined in (3.4) satisfies $\breve{G} \geq (G^* + \delta_{\hat{B}_Y})^*$.

Proof. Since J_B is coercive, lower semicontinuous, and bounded from below, $\hat{B}$ is non-empty. Since $\inf J_B = -J_B^*(0,0)$, we calculate

$$\begin{aligned}\breve{G}(y') &\geq \sup_{(\hat{x},\hat{y})\in\hat{B}} (\langle y', \hat{y}\rangle - G^*(\hat{y}) - J_B(\hat{x}, \hat{y})) - J_B^*(0,0)\\ &= \sup_{\hat{y}\in\hat{B}_Y} (\langle y', \hat{y}\rangle - G^*(\hat{y})) = (G^* + \delta_{\hat{B}_Y})^*(y'). \qquad \square\end{aligned}$$

Remark 3.6. If $\hat{B}_Y$ is convex, then $\delta_{\hat{B}_Y} = \sigma^*_{\hat{B}_Y}$ for the support function $\sigma_{\hat{B}_Y}$. As this is convex, and lower semicontinuous, we get that $\breve{G} \geq G \,\square\, \sigma_{\hat{B}_Y}$.

We always have $\breve{G} \leq G$ since $-J_B^*(0,0) \leq J_B(\bar{x}, \bar{y})$. The following establishes a lower bound on $\breve{G}$ in the our typical case of interest, with G a seminorm. It does not help interpret $\breve{G}$, but will be sufficient for developing regularisation theory in Section 5.

Lemma 3.7. *Let $F : X \to \overline{\mathbb{R}}$ be convex, proper, and lower semicontinuous, and let $G = \delta^*_{B_Y}$ be the support function of a closed convex set $B_Y \subset Y$. Pick $B \subset X \times B_Y$. Then $\breve{G}$ as defined in (3.4) satisfies $\breve{G} \geq -G(-\bullet)$.*

Proof. $(\bar{x}, \bar{y}) \in \operatorname{dom} J$ implies $\bar{y} \in B_Y$, hence $G^*(\bar{y}) = \delta_{B_Y}(\bar{y}) = 0$. Thus

$$\begin{aligned}\breve{G}(y') &= \sup_{\bar{x}\in X, \bar{y}\in Y} (\langle y', \bar{y}\rangle - J_B(\bar{x}, \bar{y})) - J_B^*(0,0)\\ &\geq \inf_{(\bar{x},\bar{y})\in B} \langle y', \bar{y}\rangle + \sup_{\tilde{x}\in X, \bar{y}\in Y} (-J_B(\tilde{x}, \bar{y})) - J_B^*(0,0)\\ &\geq \inf_{\bar{y}\in B_Y} \langle y', \bar{y}\rangle = -\delta^*_{B_Y}(-y') = -G(-y'). \qquad \square\end{aligned}$$

4 PREDICTIVE ONLINE PRIMAL-DUAL PROXIMAL SPLITTING

We now develop for (1.1) a predictive online version of the primal-dual proximal splitting (PDPS) of [11]. The structure is presented in Algorithm 4.1; our remaining work here consists of developing rules

Algorithm 4.1 Predictive online primal-dual proximal splitting (POPD)

Require: For all $k \in \mathbb{N}$, on Hilbert spaces X_k and Y_k, convex, proper, lower semicontinuous $F_{k+1} : X_{k+1} \to \overline{\mathbb{R}}$ and $G^*_{k+1}, \tilde{G}^*_{k+1} : Y_{k+1} \to \overline{\mathbb{R}}$, predictors $A_k : X_k \to X_{k+1}$ and $B_k : Y_k \to Y_{k+1}$, and $K_{k+1} \in \mathbb{L}(X_{k+1}; Y_{k+1})$. Step length parameters $\tau_{k+1}, \sigma_{k+1}, \tilde{\sigma}_{k+1} > 0$.

1: Pick initial iterates $x^0 \in X_0$ and $y^0 \in Y_0$.
2: **for** $k \in \mathbb{N}$ **do**
3: $\quad \xi^{k+1} := A_k(x^k)$ ▷ primal prediction
4: $\quad v^{k+1} := \operatorname{prox}_{\tilde{\sigma}_{k+1}\tilde{G}^*_{k+1}}(B_k(y^k) + \tilde{\sigma}_{k+1}K_{k+1}\xi^{k+1})$ ▷ dual prediction
5: $\quad x^{k+1} := \operatorname{prox}_{\tau_{k+1}F_{k+1}}(\xi^{k+1} - \tau_{k+1}K^*_{k+1}v^{k+1})$ ▷ primal step
6: $\quad y^{k+1} := \operatorname{prox}_{\sigma_{k+1}G^*_{k+1}}(v^{k+1} + \sigma_{k+1}K_{k+1}(2x^{k+1} - \xi^{k+1}))$ ▷ dual step
7: **end for**

for the step length parameters τ_{k+1}, σ_{k+1}, and $\tilde{\sigma}_{k+1}$ such that a low regret, for a suitable form of regret, is obtained. Algorithm 4.1 consists of primal and dual steps (Lines 5 and 6) that are analogous to the standard PDPS. Those are preceded by primal and dual prediction steps (Lines 3 and 4). The primal prediction is basic, based on the user-prescribed predictor A_k, but the dual prediction is somewhat more involved, imposed by a our regret theory. In particular, it involves the somewhat arbitrary functions $\tilde{G}_{k+1}$.

4.1 Assumptions and definitions

To develop the regret theory, with the general notation $u = (x, y)$, $u^k = (x^k, y^k)$, etc., we work with the following setup:

Assumption 4.1. For all $k \geq 1$, on Hilbert spaces X_k and Y_k, we assume to be given:

(i) convex, proper, and lower semicontinuous functions $F_k : X_k \to \overline{\mathbb{R}}$ and $G^*_k : Y_k \to \overline{\mathbb{R}}$, as well as $K_k \in \mathbb{L}(X_k; Y_k)$.
(ii) Primal and dual step length parameters $\tau_k, \sigma_k > 0$.
(iii) Primal and dual predictors $A_k : X_k \to X_{k+1}$ and $B_k : Y_k \to Y_{k+1}$.
(iv) Some $\tilde{\rho}_{k+1}$-strongly convex, proper, and lower semicontinuous $\tilde{G}^*_{k+1} : Y_{k+1} \to \overline{\mathbb{R}}$ and parameters $\tilde{\sigma}_{k+1} > 0$.

Further, we assume:

(v) to be given a bounded set of primal-dual comparison sequences

$$\mathcal{U} \subset \left\{ \bar{u}^{0:\infty} \in \textstyle\prod_{k=0}^{\infty} X_k \times Y_k \;\middle|\; \begin{array}{c} \bar{y}^{k+1} = \operatorname{prox}_{\tilde{\sigma}_{k+1}\tilde{G}^*_{k+1}}(\tilde{y}^{k+1} + \tilde{\sigma}_{k+1}K_{k+1}\bar{x}^{k+1}) \\ \text{for some } \tilde{y}^{k+1} =: \tilde{y}^{k+1}(\bar{u}^{k+1}) \in Y_{k+1},\ \forall k \geq 0 \end{array} \right\}$$

with which we define the set of primal comparison sequences as

$$\mathcal{B} := \{\bar{x}^{0:\infty} \mid \bar{u}^{0:\infty} \in \mathcal{U}\}.$$

(vi) for some (Lipschitz-like) factors $\Lambda_k, \Theta_k > 0$ and *prediction penalties* $\varepsilon_{k+1}, \tilde{\varepsilon}_{k+1} \in \mathbb{R}$ the primal and dual *prediction bounds*

$$\frac{1}{2}\|A_k(x^k) - \bar{x}^{k+1}\|^2_{X_{k+1}} \leq \frac{\Lambda_k}{2}\|x^k - \bar{x}^k\|^2_{X_k} + \varepsilon_{k+1} \quad \text{and} \tag{4.1a}$$

$$\frac{1}{2}\|B_k(y^k) - \tilde{y}^{k+1}\|^2_{Y_{k+1}} \leq \frac{\Theta_k}{2}\|y^k - \bar{y}^k\|^2_{Y_k} + \tilde{\varepsilon}_{k+1} \quad (\bar{u}^{0:\infty} \in \mathcal{U},\ k \in \mathbb{N}), \tag{4.1b}$$

where $\mathcal{U}$ and $\tilde{y}^{k+1}$ are as in (v), and (x^k, y^k) are generated by Algorithm 4.1.

Remark 4.2. Assumption 4.1 (v) and (vi) are not directly needed for formulating Algorithm 4.1. They are needed to develop the regret theory. The Lipschitz-like constants Λ_k and Θ_k will, however, appear in the step length rules that we develop.

In a typical case $\bar{x}^{k+1} = \bar{A}_k(\bar{x}^k)$ and $\bar{y}^{k+1} = \bar{B}_k(\bar{y}^k)$ for some true (unknown) temporal coupling operators $\bar{A}_k$ and $\bar{B}_k$ that the (known) predictors A_k and B_k approximate. Then (4.1) reads

$$\begin{aligned}\frac{1}{2}\|A_k(x^k) - \bar{A}_k(\bar{x}^k)\|^2_{X_{k+1}} &\leq \frac{\Lambda_k}{2}\|x^k - \bar{x}^k\|^2_{X_k} + \varepsilon_{k+1} \quad \text{and}\\ \frac{1}{2}\|B_k(y^k) - \bar{B}_k(\bar{y}^k)\|^2_{Y_{k+1}} &\leq \frac{\Theta_k}{2}\|y^k - \bar{y}^k\|^2_{Y_k} + \tilde{\varepsilon}_{k+1} \quad (k \in \mathbb{N})\end{aligned}$$

where the comparison points $\bar{x}^k$ and $\bar{y}^k$ are given through the recurrences $\bar{x}^{k+1} = A_k(\bar{x}^k)$ and $\bar{y}^{k+1} = \operatorname{prox}_{\tilde{\sigma}_{k+1}\tilde{G}^*_{k+1}}(B_k(\bar{y}^k) + \tilde{\sigma}_{k+1}K_{k+1}\bar{x}^{k+1})$. It may be easiest to omit the recurrences and prove the inequalities for any comparison points $\bar{x}^k$ and $\bar{y}^k$. If we had $A_k = \bar{A}_k$ and $B_k = \bar{B}_k$, and these operators were Lipschitz, we could take Λ_k and Θ_k as the corresponding Lipschitz factors and the prediction errors $\varepsilon_{k+1} = \tilde{\varepsilon}_{k+1} = 0$. Typically we would not know the true temporal coupling—or would know it only up to measurement noise—so need the prediction errors to model this lack of knowledge or noise.

Example 4.3. We can always take, and in practise take, $\tilde{G}^*_{k+1} = G^*_{k+1} + \frac{\tilde{\rho}_{k+1}}{2}\|\cdot\|^2_{Y_{k+1}}$.

We now define for all $k \geq 1$ the monotone operator[3] $H_k : X_k \times Y_k \rightrightarrows X_k \times Y_k$ and the linear preconditioned $M_k \in \mathbb{L}(X_k \times Y_k; X_k \times Y_k)$ as

$$H_k(u) := \begin{pmatrix}\partial F_k(x) + K_k^* y\\ \partial G_k^*(y) - K_k x\end{pmatrix} \quad \text{and} \quad M_k := \begin{pmatrix}\tau_k^{-1}\operatorname{Id} & -K_k^*\\ -K_k & \sigma_k^{-1}\operatorname{Id}\end{pmatrix}. \tag{4.2}$$

Then $0 \in H_k(\hat{u}^k)$ encodes the primal-dual optimality conditions (3.1) for the static problem (1.6) while Algorithm 4.1 can be written in implicit form as

$$0 \in H_k(u^k) + M_k(u^k - z^k) \quad (k \geq 1) \tag{4.3}$$

for

$$z^{k+1} := (\xi^{k+1}, v^{k+1}) := S_k(u^k), \quad S_k(u) := \begin{pmatrix}A_k(x)\\ \operatorname{prox}_{\tilde{\sigma}_{k+1}\tilde{G}^*_{k+1}}(B_k(y) + \tilde{\sigma}_{k+1}K_{k+1}A_k(x))\end{pmatrix} \quad (k \geq 0). \tag{4.4}$$

We now derive regret estimates based on the partial primal gaps of Section 3.

4.2 A general regret estimate

We need the following *strong non-expansivity* from the dual predictor. The result is standard, but difficult to find explicitly stated in the literature for $\gamma > 0$:

Lemma 4.4. *On a Hilbert space X, suppose $F : X \to \overline{\mathbb{R}}$ is convex, proper, and γ-strongly subdifferentiable. Then prox_F is $(1+\gamma)$-strongly non-expansive:*

$$(1+\gamma)\|\operatorname{prox}_F(x) - \operatorname{prox}_F(\tilde{x})\|^2_X \leq \langle \operatorname{prox}_F(x) - \operatorname{prox}_F(\tilde{x}), x - \tilde{x}\rangle \quad (x, \tilde{x} \in X).$$

Proof. Let $y := \operatorname{prox}_F(x)$. By definition, $y + q = x$ and $\tilde{y} + \tilde{q} = \tilde{x}$ for some $q \in \partial F(y)$ and $\tilde{q} \in \partial F(\tilde{y})$. Since ∂F is γ-strongly monotone, $\langle q - \tilde{q}, y - \tilde{y}\rangle \geq \gamma\|y - \tilde{y}\|^2$. Thus

$$(1+\gamma)\|y - \tilde{y}\|^2 = \langle y - \tilde{y}, x - \tilde{x} - (q - \tilde{q})\rangle + \gamma\|y - \tilde{y}\|^2 \leq \langle y - \tilde{y}, x - \tilde{x}\rangle. \qquad \square$$

[3] The double arrow signifies that the map is set-valued.

The next lemma derives basic step length conditions, which we will further develop in Section 4.3, from basic properties of the linear preconditioner M_k and an overall primal-dual prediction bound analogous to (4.1). The "testing" parameters $\varphi_k, \psi_k, \eta_k > 0$ model the respective primal, dual, and joint (e.g., gap) convergence or regret rates. They are coupled via (4.5a) to the step length parameters. Any one of these parameters is superfluous given the others, but all are included for notational and conceptual convenience. The testing parameters are not directly required in Algorithm 2.1, but will serve to study "regret rates".

Lemma 4.5. *Suppose Assumption 4.1 holds. Fix $k \in \mathbb{N}$ and assume for some $\kappa \in (0,1)$ and* testing parameters $\eta_k, \varphi_k, \psi_k > 0$, *the step length conditions*

$$\eta_k = \varphi_k \tau_k = \psi_k \sigma_k, \tag{4.5a}$$ *(primal-dual coupling)*

$$\tilde{\rho}_{k+1} \geq \frac{\Theta_k \eta_{k+1} \tilde{\sigma}_{k+1}^{-2}}{2\kappa(1+\sigma_k\rho_k)\psi_k} + \frac{1}{2\sigma_{k+1}} - \tilde{\sigma}_{k+1}^{-1}, \tag{4.5b}$$ *(proximal predictor restriction)*

$$\varphi_k(1+\gamma_k\tau_k) \geq \varphi_{k+1}\Lambda_k + \frac{\varphi_k\tau_k\sigma_k\|K_k\|^2}{(1-\kappa)(1+\sigma_k\rho_k)}, \tag{4.5c}$$ *(primal metric update) and*

$$1 \geq \tau_k\sigma_k\|K_k\|^2 \tag{4.5d}$$ *(metric positivity).*

Let

$$\Gamma_k := \eta_k \begin{pmatrix} \gamma_k \operatorname{Id} & 2K_k^* \\ -2K_k & \rho_k \operatorname{Id} \end{pmatrix}. \tag{4.6}$$

Then $\eta_k M_k$ is self-adjoint and positive semidefinite, $\eta_k M_k + \Gamma_k$ is positive semidefinite, and we have the overall prediction bound

$$\frac{1}{2}\|z^{k+1} - \bar{x}^{k+1}\|^2_{\eta_{k+1}M_{k+1}} \leq \frac{1}{2}\|x^k - \bar{x}^k\|^2_{\eta_k M_k + \Gamma_k} + \varphi_{k+1}\varepsilon_{k+1} + \frac{\kappa(1+\sigma_k\rho_k)\psi_k}{2\Theta_k}\tilde{\varepsilon}_{k+1} \quad (k = 0, \dots, N-1). \tag{4.7}$$

Proof. Using (4.5a) and Young's inequality, we expand and estimate

$$\eta_k M_k = \begin{pmatrix} \varphi_k \operatorname{Id} & -\eta_k K_k^* \\ -\eta_k K_k & \psi_k \operatorname{Id} \end{pmatrix} \geq \begin{pmatrix} \varphi_k \operatorname{Id} - \eta_k^2 \psi_k^{-1} K_k^* K_k & 0 \\ 0 & 0 \end{pmatrix}. \tag{4.8}$$

Thus $\eta_k M_k$ is self-adjoint due to (4.5a) and positive semidefinite due to (4.5d) and (4.5a). It follows, using Young's inequality, that

$$\eta_k M_k + \Gamma_k \simeq \begin{pmatrix} \varphi_k(1+\gamma_k\tau_k)\operatorname{Id} & -\eta_k K_k^* \\ -\eta_k K_k & \psi_k(1+\rho_k\sigma_k)\operatorname{Id} \end{pmatrix} \geq \begin{pmatrix} \varphi_k(1+\gamma_k\tau_k)\operatorname{Id} - \frac{\eta_k^2}{\psi_k(1+\rho_k\sigma_k)} K_k^* K_k & 0 \\ 0 & 0 \end{pmatrix}. \tag{4.9}$$

Thus $\eta_k M_k + \Gamma_k$ is positive semidefinite by (4.5c) and (4.5a).

We still need to prove (4.7). Writing $(\xi^{k+1}, v^{k+1}) := z^{k+1} = S_k(u^k)$, we have

$$\begin{aligned} \frac{1}{2}\|z^{k+1} - \bar{u}^{k+1}\|^2_{\eta_{k+1}M_{k+1}} = {} & \frac{\varphi_{k+1}}{2}\|A_k(x^k) - \bar{x}^{k+1}\|^2 + \frac{\psi_{k+1}}{2}\|v^{k+1} - \bar{y}^{k+1}\|^2 \\ & - \eta_{k+1}\langle K_{k+1}(\xi^{k+1} - \bar{x}^{k+1}), v^{k+1} - \bar{y}^{k+1}\rangle \end{aligned} \tag{4.10}$$

as well as

$$\begin{aligned} \frac{1}{2}\|u^k - \bar{u}^k\|^2_{\eta_k M_k + \Gamma_k} = {} & \frac{\varphi_k(1+\gamma_k\tau_k)}{2}\|x^k - \bar{x}^k\|^2 + \frac{\psi_k(1+\rho_k\sigma_k)}{2}\|y^k - \bar{y}^k\|^2 \\ & - \eta_k\langle K_k(x^k - \bar{x}^k), y^k - \bar{y}^k\rangle. \end{aligned} \tag{4.11}$$

Since $\tilde{G}_{k+1}$ is ($\tilde{\rho}_{k+1}$-strongly) convex, by Lemma 4.4, (4.4), and Assumption 4.1, (v)

$$(1+\tilde{\sigma}_{k+1}\tilde{\rho}_{k+1})\|v^{k+1}-\bar{y}^{k+1}\|^2 \le \langle v^{k+1}-\bar{y}^{k+1}, B_k(y^k)-\tilde{y}^{k+1}+\tilde{\sigma}_{k+1}K_{k+1}(\xi^{k+1}-\bar{x}^{k+1})\rangle.$$

By (4.5a) and (4.5b),

$$-\eta_{k+1}\tilde{\sigma}_{k+1}^{-1}(1+\tilde{\sigma}_{k+1}\tilde{\rho}_{k+1})+\frac{\Theta_k\eta_{k+1}^2\tilde{\sigma}_{k+1}^{-2}}{2\kappa(1+\sigma_k\rho_k)\psi_k} \le -\frac{\psi_{k+1}}{2}.$$

Consequently, also using (4.5a), (4.1b), and Young's inequality, we obtain

$$\begin{aligned}
&-\eta_{k+1}\langle K_{k+1}(\xi^{k+1}-\bar{x}^{k+1}), v^{k+1}-\bar{y}^{k+1}\rangle\\
&\quad= -\eta_{k+1}\tilde{\sigma}_{k+1}^{-1}\langle B_k(y^k)-\tilde{y}^{k+1}+\tilde{\sigma}_{k+1}K_{k+1}(\xi^{k+1}-\bar{x}^{k+1}), v^{k+1}-\bar{y}^{k+1}\rangle\\
&\qquad+\eta_{k+1}\tilde{\sigma}_{k+1}^{-1}\langle B_k(y^k)-\tilde{y}^{k+1}, v^{k+1}-\bar{y}^{k+1}\rangle\\
&\quad\le -\eta_{k+1}\tilde{\sigma}_{k+1}^{-1}(1+\tilde{\sigma}_{k+1}\tilde{\rho}_{k+1})\|v^{k+1}-\bar{y}^{k+1}\|^2+\eta_{k+1}\tilde{\sigma}_{k+1}^{-1}\langle B_k(y^k)-\tilde{y}^{k+1}, v^{k+1}-\bar{y}^{k+1}\rangle\\
&\quad\le -\frac{\psi_{k+1}}{2}\|v^{k+1}-\bar{y}^{k+1}\|^2+\frac{\kappa(1+\sigma_k\rho_k)\psi_k}{2\Theta_k}\|B_k(y^k)-\tilde{y}^{k+1}\|^2.\\
&\quad\le -\frac{\psi_{k+1}}{2}\|v^{k+1}-\bar{y}^{k+1}\|^2+\frac{\kappa(1+\sigma_k\rho_k)\psi_k}{2\Theta_k}\left(\frac{\Theta_k}{2}\|y^k-\bar{y}^k\|^2+\tilde{\varepsilon}_{k+1}\right).
\end{aligned}$$

Applying this and (4.1a) in (4.10), we obtain for $p_{k+1} := \varphi_{k+1}\varepsilon_{k+1}+\frac{\kappa(1+\sigma_k\rho_k)\psi_k}{2\Theta_k}\tilde{\varepsilon}_{k+1}$ that

$$\frac{1}{2}\|z^{k+1}-\bar{u}^{k+1}\|^2_{\eta_{k+1}M_{k+1}} \le \frac{\varphi_{k+1}\Lambda_k}{2}\|x^k-\bar{x}^k\|^2+\frac{\kappa(1+\sigma_k\rho_k)\psi_k}{2}\|y^k-\bar{y}^k\|^2+p_{k+1}. \tag{4.12}$$

We also have by Young's inequality

$$\begin{aligned}
&\eta_k\langle K_k(x^k-\bar{x}^k), y^k-\bar{y}^k\rangle\\
&\qquad\le \frac{\eta_k^2}{2(1-\kappa)(1+\sigma_k\rho_k)\psi_k}\|K_k(x^k-\bar{x}^k)\|^2+\frac{(1-\kappa)(1+\sigma_k\rho_k)\psi_k}{2}\|y^k-\bar{y}^k\|^2.
\end{aligned}$$

Hence (4.11) gives

$$\begin{aligned}
-\frac{1}{2}\|u^k-\bar{u}^k\|^2_{\eta_kM_k+\Gamma_k} \le &\left(\frac{\eta_k^2\|K_k\|^2}{2(1-\kappa)(1+\sigma_k\rho_k)\psi_k}-\frac{\varphi_k(1+\gamma_k\tau_k)}{2}\right)\|x^k-\bar{x}^k\|^2\\
&-\frac{\kappa(1+\sigma_k\rho_k)\psi_k}{2}\|y^k-\bar{y}^k\|^2.
\end{aligned} \tag{4.13}$$

Combined, (4.12) and (4.13) show that

$$\begin{aligned}
\frac{1}{2}\|z^{k+1}-\bar{u}^{k+1}\|^2_{\eta_{k+1}M_{k+1}}-\frac{1}{2}\|u^k-\bar{u}^k\|^2_{\eta_kM_k+\Gamma_k} &\le \varphi_{k+1}\varepsilon_{k+1}+p_{k+1}\\
&+\left(\frac{\varphi_{k+1}\Lambda_k}{2}+\frac{\eta_k^2\|K_k\|^2}{2(1-\kappa)(1+\sigma_k\rho_k)\psi_k}-\frac{\varphi_k(1+\gamma_k\tau_k)}{2}\right)\|x^k-\bar{x}^k\|^2.
\end{aligned}$$

From here (4.5c) shows (4.7). □

To state the final regret estimate, for brevity we define

$$F_{1:N}(x^{1:N}) := \sum_{k=0}^{N-1}\eta_kF_{k+1}(x^{k+1}),\quad G_{1:N}(y^{1:N}) := \sum_{k=0}^{N-1}\eta_kG_{k+1}(\eta_k^{-1}y^{k+1}),\quad \text{and}$$
$$K_{1:N}x^{1:N} := (\eta_0K_1x^1,\dots,\eta_{N-1}K_Nx^N).$$

We recall the comparison sets $\mathcal{U}$ and $\mathcal{B}$ and from Assumption 4.1 and the slicing notation $\mathcal{U}_{n:m}$ and $\mathcal{B}_{n:m}$ form Section 1. With these we also define

$$\check{G}_{1:N}(y'_{1:N}) = \sup_{\bar{x}^{1:N},\bar{y}^{1:N}} \Big(\langle y'_{1:N}, \bar{y}^{1:N} \rangle - G^*_{1:N}(\bar{y}^{1:N}) - J_{\mathcal{U}_{1:N}}(\bar{x}^{1:N}, \bar{y}^{1:N}) \Big) - J^*_{\mathcal{U}_{1:N}}(0,0) \tag{4.14}$$

with the supremum running over $\bar{x}^{1:N} \in X_1 \times \cdots \times X_N$ and $\bar{y}^{1:N} \in Y_1 \times \cdots \times Y_N$ and

$$J_{\mathcal{U}_{1:N}}(\tilde{x}^{1:N}, \tilde{y}^{1:N}) := [F_{1:N} + G_{1:N} \circ K_{1:N}](\tilde{x}^{1:N}) + \delta_{\mathcal{U}_{1:N}}(\tilde{x}^{1:N}, \tilde{y}^{1:N}).$$

Observe that $G^*_{1:N}(y^{1:N}) = \sum_{k=0}^{N-1} \eta_k G^*_{k+1}(y^{k+1})$ and

$$[F_{1:N} + G_{1:N} \circ K_{1:N}](x^{1:N}) = \sum_{k=0}^{N-1} \eta_k [F_{k+1} + G_{k+1} \circ K_{k+1}](x^{k+1}). \tag{4.15}$$

After the next main regret estimate, we comment upon its assumptions and claim.

Theorem 4.6. *Suppose Assumption 4.1 and the step length bounds (4.5) hold for $u^{1:N}$ generated by Algorithm 4.1 for an initial $u^0 \in X_0 \times Y_0$. Then*

$$[F_{1:N} + \check{G}_{1:N} \circ K_{1:N}](x^{1:N}) - \inf_{\bar{x}^{1:N} \in \mathcal{B}_{1:N}} [F_{1:N} + G_{1:N} \circ K_{1:N}](\bar{x}^{1:N}) + \sum_{k=0}^{N-1} \frac{\|u^{k+1} - S_k(u^k)\|^2_{\eta_{k+1}M_{k+1}}}{2}$$
$$\leq e_N := \sup_{\bar{u}^0 \in \mathcal{U}_0} \frac{1}{2} [\![u^0 - \bar{u}^0]\!]^2_{\eta_0 M_0 + \Gamma_0} + \sum_{k=0}^{N-1} \left(\varepsilon_{k+1}\varphi_{k+1} + \frac{\kappa(1+\sigma_k\rho_k)\psi_k}{2\Theta_k} \tilde{\varepsilon}_{k+1} \right).$$

Proof. For brevity, and to not abuse norm notation when Γ_k is not positive semi-definite, we write $[\![x]\!]^2_{\Gamma_k} := \langle x, x \rangle_{\Gamma_k}$. By Lemma 4.5, $\eta_k M_k$ and $\eta_k M_k + \Gamma_k$ are positive semi-definite, so we may use the norm notation with them. For H_k defined (4.2) and Γ_k and η_k in (4.6), the (strong) convexity of F_k and G^*_k yield

$$\langle H_k(u^k), u^k - \bar{u}^k \rangle_{\eta_k} \geq \frac{1}{2} [\![u^k - \bar{u}^k]\!]^2_{\Gamma_k} + \mathcal{G}^H_k \quad (k = 1, \ldots, N) \tag{4.16}$$

for

$$\begin{aligned} \mathcal{G}^H_{k+1} := \eta_k [& F_{k+1}(x^{k+1}) - F_{k+1}(\bar{u}^{k+1}) + G^*_{k+1}(y^{k+1}) - G^*_{k+1}(\bar{y}^{k+1}) \\ & - \langle K^*_{k+1} y^{k+1}, \bar{u}^{k+1} \rangle + \langle K_{k+1} x^{k+1}, \bar{y}^{k+1} \rangle]. \end{aligned} \tag{4.17}$$

Following the testing methodology of [30, 14], we pick any $\bar{u}^k \in X_k \times Y_k$ and apply the linear "testing operator" $\langle \,\cdot\,, u^k - \bar{u}^k \rangle_{\eta_k}$ to both sides of (4.3). This followed by (4.16) yields

$$0 \geq \langle u^k - z^k, u^k - \bar{u}^k \rangle_{\eta_k M_k} + \frac{1}{2} [\![u^k - \bar{u}^k]\!]^2_{\Gamma_k} + \mathcal{G}^H_k \quad (k = 1, \ldots, N).$$

Pythagoras' identity (2.8) for the inner product and norm with respect to the operator $\eta_k M_k$ now yields

$$\frac{1}{2} \|z^k - \bar{u}^k\|^2_{\eta_k M_k} \geq \frac{1}{2} \|u^k - \bar{u}^k\|^2_{\eta_k M_k + \Gamma_k} + \mathcal{G}^H_k + \frac{1}{2} \|u^k - z^k\|^2_{\eta_k M_k} \quad (k = 1, \ldots, N).$$

We now take $\bar{u}^{0:N} \in \mathcal{U}_{0:N}$ and apply the prediction bound (4.7) from Lemma 4.5 to obtain

$$\frac{1}{2} \|u^k - \bar{u}^k\|^2_{\eta_k M_k + \Gamma_k} + \left(\varphi_{k+1}\varepsilon_{k+1} + \frac{\kappa(1+\sigma_k\rho_k)\psi_k}{2\Theta_k} \tilde{\varepsilon}_{k+1} \right)$$
$$\geq \frac{1}{2} \|u^{k+1} - \bar{u}^{k+1}\|^2_{\eta_{k+1} M_{k+1} + \Gamma_{k+1}} + \mathcal{G}^H_{k+1} + \frac{1}{2} \|u^k - z^k\|^2_{\eta_k M_k} \quad (k = 1, \ldots, N-1).$$

Summing over such k and taking the supremum over $\bar{u}^{0:N} \in \mathcal{U}_{0:N}$, we get

$$\sup_{\bar{u}^{0:N}\in\mathcal{U}_{0:N}} \sum_{k=0}^{N-1} \left(\mathcal{G}_{k+1}^H + \frac{1}{2}\|u^{k+1} - z^{k+1}\|^2_{\eta_{k+1}M_{k+1}} \right) \leq e_N.$$

By Lemma 3.1 applied to $K = K_{1:N}$, $F = F_{1:N}$ and $G^* = G^*_{1:N}$ and (4.17) we obtain

$$\sup_{\bar{u}^{1:N}\in\mathcal{U}_{1:N}} \sum_{k=0}^{N-1} \mathcal{G}_{k+1}^H \geq [F_{1:N} + \check{G}_{1:N} \circ K_{1:N}](x^{1:N}) - \inf_{\bar{x}^{1:N}\in\mathcal{B}_{1:N}} [F_{1:N} + G_{1:N} \circ K_{1:N}](\bar{x}^{1:N}).$$

Since $z^{k+1} := S_k(u^k)$ by (4.4), these two inequalities together verify the claim. □

Remark 4.7 (Satisfying the conditions). Assumption 4.1 is structural. Aside from $\tilde{G}_{k+1}$, everything in it depends on the application problem and the predictors we can design for it. The function $\tilde{G}_{k+1}$ can be taken as in Example 4.3. The step length bounds (4.5) can be satisfied via the choices in the next Section 4.3.

Remark 4.8 (Interpretation of the dual comparison sequence). Let $\tilde{y}^{k+1} = \bar{B}_k(\bar{y}^k)$ for a dual temporal coupling operator $\bar{B}_k$. Then the definition of $\mathcal{U}$ in Assumption 4.1 (v) updates the dual comparison variable as

$$\bar{y}^{k+1} := \operatorname{prox}_{\tilde{\sigma}_{k+1}\tilde{G}^*_{k+1}}(\bar{B}_k(\bar{y}^k) + \tilde{\sigma}K_{k+1}\bar{x}^{k+1}) \tag{4.18}$$

This amounts to the POFB of Section 2 applied with the predictor $\bar{B}_k$ and the step length parameter $\tau_{k+1} = \tilde{\sigma}_{k+1}$ to the formal problem

$$\min_{y^1,y^2,\ldots} \sum_{k=1}^{\infty} \tilde{G}^*_k(y^k) - \langle K_k\bar{x}^k, y^k\rangle, \quad y^{k+1} = \bar{B}_k(y^k)$$

An "optimal" $\hat{y}^k$, achieving $\inf_y \tilde{G}^*_k(y) - \langle K_k\bar{x}^k, y\rangle$, would give

$$[F_k + \tilde{G}_k \circ K_k](\bar{x}^k) = F_k(\bar{x}^k) + \langle K_k\bar{x}^k, \hat{y}^k\rangle - \tilde{G}^*_k(\hat{y}^k).$$

This is approximated by $\bar{y}^{k+1}$ generated by (4.18), better as $\tilde{\sigma}_k \nearrow \infty$. In the setting of Example 4.3, if also $\tilde{\rho}_k \searrow 0$, then we get closer to calculating $[F_k + G_k \circ K_k](\bar{x}^k)$.

4.3 SPECIFIC STEP LENGTH CHOICES

We now develop explicit step length rules that satisfy the step length conditions (4.5), and then interpret Theorem 4.6 for them. The proof of the next lemma is immediate:

Lemma 4.9. *The right hand side of* (4.5b) *is minimised by* $\tilde{\sigma}_{k+1} = \frac{\Theta_k\eta_{k+1}}{\kappa(1+\sigma_k\rho_k)\psi_k}$. *With this choice* (4.5b) *reads* $2\eta_{k+1}\tilde{\rho}_{k+1} \geq \psi_{k+1} - \kappa\Theta_k^{-1}(1+\sigma_k\rho_k)\psi_k$.

The following examples use Lemma 4.9:

Example 4.10 (Constant step length and testing parameters). In Algorithm 4.1, take as the step length parameters $\tau_k \equiv \tau$, $\sigma_k \equiv \sigma$, and $\tilde{\sigma}_{k+1} = \frac{\Theta_k\sigma}{\kappa(1+\sigma\rho_k)}$ for some constant $\tau, \sigma > 0$ and $\kappa \in (0,1)$ satisfying for the strong convexity factors γ_k, ρ_k, $\tilde{\rho}_{k+1}$ and the Lipschitz-like factors Θ_k, Λ_k from Assumption 4.1 the inequalities

$$\tilde{\rho}_{k+1} \geq \frac{1}{2\sigma}\left(1 - \frac{\kappa(1+\sigma\rho_k)}{\Theta_k}\right), \quad 1 + \gamma_k\tau \geq \Lambda_k + \frac{\tau\sigma\|K_k\|^2}{(1-\kappa)(1+\sigma\rho_k)}, \quad \text{and} \quad 1 \geq \tau\sigma\|K_k\|^2. \tag{4.19}$$

(By Example 4.3, we may simply define $\tilde{\rho}_{k+1}$ through the first expression if we choose to take

$\tilde{G}^*_{k+1} = G^*_{k+1} + \frac{\tilde{\rho}_{k+1}}{2}\|\cdot\|^2_{Y_{k+1}}$.) Then (4.5) holds for the testing parameters $\eta_k \equiv \tau$, $\varphi_k \equiv 1$, and $\psi_k \equiv \frac{\tau}{\sigma}$. In this case, Theorem 4.6 shows for an initialisation-dependent constant C_0 that

$$[F_{1:N} + \check{G}_{1:N} \circ K_{1:N}](x^{1:N}) - \inf_{\bar{x}^{1:N} \in \mathcal{B}_{1:N}} [F_{1:N} + G_{1:N} \circ K_{1:N}](\bar{x}^{1:N}) \leq C_0 + \sum_{k=0}^{N-1} \left(\varepsilon_{k+1} + \frac{\kappa(1+\sigma\rho_k)\tau}{2\Theta_k\sigma} \tilde{\varepsilon}_{k+1} \right).$$

Suppose $\sup_k \rho_k \leq \overline{\rho}$ and $\inf_k \Theta_k \geq \underline{\Theta}$ for some $\overline{\rho}, \underline{\Theta} > 0$ (such as when ρ_k and Θ_k are constant in k). Minding the sum expression (4.15), where now $\eta_k = \tau$, for a constant $C > 0$, we get

$$\frac{1}{N}[F_{1:N} + \check{G}_{1:N} \circ K_{1:N}](x^{1:N}) - \inf_{\bar{x}^{1:N} \in \mathcal{B}_{1:N}} \frac{\tau}{N} \sum_{k=0}^{N-1} [F_{k+1} + G_{k+1} \circ K_{k+1}](\bar{x}^{k+1}) \leq \frac{C_0 + C\sum_{k=0}^{N-1}(\varepsilon_{k+1} + \tilde{\varepsilon}_{k+1})}{N}.$$

Exact interpretation requires being able to calculate $\check{G}_{1:N}$, however we can make a rough interpretation. We distinguish two cases:

(a) If $\sum_{k=0}^{\infty}(\varepsilon_{k+1} + \tilde{\varepsilon}_{k+1}) < \infty$, then the left hand side converges below zero as $N \to \infty$. Roughly, subject to how well we can measure with $\check{G}_{1:N}$ in place of $G_{1:N}$, this says that asymptotically $x^{1:N}$ are at least as good solutions of the averaged problem $\inf_{x^{1:N}} \frac{\tau}{N}\sum_{k=0}^{N-1}[F_{k+1} + G_{k+1} \circ K_{k+1}](x^{k+1})$ as the best constrained $\bar{x}^{1:N} \in \mathcal{B}_{1:N}$.

(b) If $\frac{1}{N}\sum_{k=0}^{N-1}(\varepsilon_{k+1} + \tilde{\varepsilon}_{k+1}) \leq \delta$ for some constant $\delta > 0$, then, again subject to how well we can measure with $\check{G}_{1:N}$ in place of $G_{1:N}$, this says that asymptotically $x^{1:N}$ stays "within average noise level" δ of the best $\bar{x}^{1:N} \in \mathcal{B}_{1:N}$.

The bounds on the the prediction errors ε_{k+1} and $\tilde{\varepsilon}_{k+1}$ can be interpreted as the noise level of the "measurements" A_k and B_k of the true temporal coupling operators $\bar{A}_k$ and $\bar{B}_k$ either vanishing or staying bounded (on average). In the optical flow example, to be further studied in Section 5, this means that the noise level of the displacement field measurements has to vanish or stay bounded (on average).

Example 4.11 (Everything constant). In particular, in Example 4.10, if the strong convexity and Lipshitz-like parameters are constant, $\tilde{\rho}_{k+1} \equiv \tilde{\rho}$, $\gamma_k \equiv \gamma$, and $\Theta_k \equiv \Theta$, and $\Lambda_k \equiv \Lambda$, with no dual strong convexity, $\rho_k = 0$, and we take $\tilde{\sigma}_{k+1} \equiv \tilde{\sigma} = \frac{\Theta\sigma}{\kappa}$, then (4.19), hence (4.5), hold if

$$\tilde{\rho} \geq \frac{1 - \kappa\Theta^{-1}}{2\sigma}, \quad 1 + \gamma\tau \geq \Lambda + \frac{\tau\sigma\|K_k\|^2}{1-\kappa}, \quad \text{and} \quad 1 \geq \tau\sigma\|K_k\|^2.$$

Examples 4.10 and 4.11 give no growth for the testing parameters φ_k, ψ_k, and η_k. We now look at one case when this is possible and what happens then.

Example 4.12 (Exponential testing parameters with constant step lengths). In Algorithm 4.1, take $\tau_k \equiv \tau$, $\sigma_{k+1} \equiv \sigma$, as well as $\tilde{\sigma}_{k+1} = \kappa^{-1}\Theta_k\sigma$ for some constant $\tau, \sigma > 0$ satisfying for the strong convexity factors $\gamma_k, \rho_k, \tilde{\rho}_{k+1}$ and the Lipschitz-like factors Θ_k, Λ_k from Assumption 4.1, for some $\kappa \in (0,1)$ the inequalities

$$\tilde{\rho}_{k+1} \geq \frac{1 - \kappa\Theta_k^{-1}}{2\sigma}, \quad 1 + \gamma_k\tau \geq \frac{\tau\sigma\|K_k\|^2}{(1-\kappa)(1+\rho_k\sigma)} + (1+\rho_k\sigma)\Lambda_k, \quad \text{and} \quad 1 \geq \tau\sigma\|K_k\|^2.$$

Then (4.5) holds with $\eta_k = \varphi_k\tau$, $\varphi_{k+1} = \varphi_k(1+\rho_k\sigma)$, and $\psi_k = \frac{\tau}{\sigma}\varphi_k$. In this case Theorem 4.6 shows

for some initialisation-dependent constant C_0 that

$$[F_{1:N} + \breve{G}_{1:N} \circ K_{1:N}](x^{1:N}) - \inf_{\bar{x}^{1:N} \in \mathcal{B}_{1:N}} [F_{1:N} + G_{1:N} \circ K_{1:N}](\bar{x}^{1:N}) \\ \leq C_0 + \sum_{k=0}^{N-1} \varphi_k \left(\varepsilon_{k+1}(1 + \gamma_k \tau) + \frac{\kappa(1 + \sigma\rho_k)\tau}{2\Theta_k \sigma} \tilde{\varepsilon}_{k+1} \right).$$

Suppose for simplicity that $\sup_k \rho_k \leq \overline{\rho}$, $\sup_k \gamma_k \leq \overline{\gamma}$, and $\inf_k \Theta_k \geq \underline{\Theta}$ for some $\overline{\rho}, \overline{\gamma}, \underline{\Theta} > 0$ (such as when ρ_k, γ_k, and Θ_k are constant in k). Then, minding the sum expression (4.15), where now $\eta_k = \tau\varphi_k$, this gives for a constant $C > 0$ the result

$$\frac{[F_{1:N} + \breve{G}_{1:N} \circ K_{1:N}](x^{1:N})}{\tau \sum_{k=0}^{N-1} \varphi_k} - \inf_{\bar{x}^{1:N} \in \mathcal{B}_{1:N}} \frac{\sum_{k=0}^{N-1} \varphi_k [F_{k+1} + G_{k+1} \circ K_{k+1}](\bar{x}^{k+1})}{\sum_{k=0}^{N-1} \varphi_k} \\ \leq C_0 + C \frac{\sum_{k=0}^{N-1} \varphi_k \left(\varepsilon_{k+1} + \tilde{\varepsilon}_{k+1} \right)}{\sum_{k=0}^{N-1} \varphi_k}.$$

Exact interpretation requires being able to calculate $\breve{G}_{1:N}$, however, as in Example 4.10, we can roughly interpret two cases:

(a) If $\lim_{N\to\infty} \sum_{k=0}^{N-1} \varphi_k \left(\varepsilon_{k+1} + \tilde{\varepsilon}_{k+1} \right) / \sum_{k=0}^{N-1} \varphi_k = 0$, the left hand side converges below zero as $N \to \infty$. Roughly, subject to how well we can measure with $\breve{G}_{1:N}$ in place of $G_{1:N}$, this says that asymptotically $x^{1:N}$ are at least as good solutions of the *weighted*-averaged problem $\inf_{x^{1:N}} \frac{1}{\sum_{k=0}^{N-1} \varphi_k} \sum_{k=0}^{N-1} \varphi_k [F_{k+1} + G_{k+1} \circ K_{k+1}](x^{k+1})$ as the best constrained $\bar{x}^{1:N} \in \mathcal{B}_{1:N}$.

(b) If $\sup_N \sum_{k=0}^{N-1} \varphi_k \left(\varepsilon_{k+1} + \tilde{\varepsilon}_{k+1} \right) / \sum_{k=0}^{N-1} \varphi_k \leq \delta$ for a constant δ, then, subject to how well we can measure with $\breve{G}_{1:N}$ in place of $G_{1:N}$, this says that asymptotically $x^{1:N}$ stay "within *weighted*-average noise level" δ of the best $\bar{x}^{1:N} \in \mathcal{B}_{1:N}$.

Since $\varphi_{k+1} = \varphi_k(1 + \rho_k\sigma)$ is increasing, later iterates are weighted more. If $\inf_k \rho_k > 0$, then φ_k grows exponentially, so the later iterates have exponentially more importance. Thus we can make worse measurements of the early data frames without significantly affecting the quality of the later iterates. If ε_{k+1} and $\tilde{\varepsilon}_{k+1}$ are noise levels of the measurements A_k and B_k of some true temporal coupling operators $\bar{A}_k$ and $\bar{B}_k$, the noise levels have to converge to zero for (a) or stay bounded for (b).

5 OPTICAL FLOW

We now apply the previous sections to optical flow. For numerical accuracy, we use the more fundamental displacement field model instead of the linearised PDE model (transport equation). For simplicity, and to keep the static problems convex, we concentrate on constant-in-space (but not time) displacement fields. This makes our work applicable to computational image stabilisation (shake reduction) in still or video cameras, compare [28, 37], based on rapid successions of very noisy images. We start in Section 5.1 with a known displacement field—as could be estimated using acceleration sensors on cameras. Afterwards in Section 5.2 we include the estimation of the displacement field into our model.

5.1 KNOWN DISPLACEMENT FIELD

Denoting by $\delta > 0$ the noise level, we start by assuming to be given in each frame, i.e., on each iteration, a noisy measurement $b_\delta^k \in X$ of a true image $\bar{b}^k \in X$ and a noisy measurement $v_\delta^k \in V$ of a true displacement field $\bar{v}^k \in V$. We assume the measured displacement fields v_δ^k bijective. The finite-dimensional subspaces $X \subset L^2(\Omega)$, $Y \subset L^2(\Omega;\mathbb{R}^2)$, and $V \subset L^2(\Omega;\Omega) \cap C^2(\Omega;\Omega)$ on a domain $\Omega \subset \mathbb{R}^2$ we equip with the L^2-norm. We write (1.3) in min-max form with

$$F_k^\delta(x) := \frac{1}{2}\|b_\delta^k - x\|_X^2, \quad (G_k^\alpha)^*(y) := \delta_{\alpha B}(y), \quad \text{and} \quad K_k = D, \tag{5.1a}$$

for B the product of pointwise unit balls and $D : X \to L^2(\Omega)$ a discretised differential operator. For the primal and dual predictors we take

$$A_k^\delta(x) := x \circ v_\delta^k \quad \text{and} \quad B_k^\delta(y) := y \circ v_\delta^k, \tag{5.1b}$$

In the dual predictor of the POPD, we take $\tilde{G}_k^* = (G_k^\alpha)^* + \frac{\tilde{\rho}_k}{2}\|\cdot\|_{L^2(\Omega)}^2$ following Example 4.3. Thus $\tilde{G}_k^*$ is the Fenchel conjugate of the Huber/Moreau–Yosida-regularised 1-norm.

REGARDING THE REGRET AND REGULARISATION THEORY

Let the true displacement fields $\bar{v}^k \in H^1(\mathbb{R}^2;\mathbb{R}^2)$, $(k \in \mathbb{N})$, and let $\mathcal{U}_0 \subset X \times Y$ be bounded. To satisfy Assumption 4.1 (v), we take for some $M > 0$,

$$\mathcal{U} := \left\{ \bar{u}^{0:\infty} \;\middle|\; \begin{array}{l} \bar{u}^0 \in \mathcal{U}_0,\ \bar{x}^{k+1} = \bar{x}^k \circ \bar{v}^k,\ \bar{x}^k \in H^1(\mathbb{R}^2),\ \bar{y}^k \in H^1(\mathbb{R}^2;\mathbb{R}^2),\ \|\nabla \bar{y}^k\|_{2,\infty}^2 \le M, \\ \|\nabla \bar{x}^k\|_{2,\infty}^2 \le M,\ \bar{y}^{k+1} = \operatorname{prox}_{\tilde{\sigma}_{k+1}\tilde{G}_{k+1}^*}(\bar{y}^k \circ \bar{v}^k + \tilde{\sigma}_{k+1}K_{k+1}\bar{x}^{k+1}),\ \forall k \ge 0 \end{array} \right\} \tag{5.2a}$$

as the comparison set. With a slight abuse of notation we also write $\mathcal{U}$ for the corresponding set with the domain of each $\bar{u}^k$ restricted to Ω. We assume that the ground-truth images

$$\bar{b}^{0:\infty} \in \mathcal{B} := \{\bar{x}^{0:\infty} \mid \bar{u}^{0:\infty} \in \mathcal{U}\}. \tag{5.2b}$$

Because the iterates y^k are in a finite-dimensional subspace, bounding $\|\nabla \bar{y}^k\|_{2,\infty}^2$ is no difficulty.

To satisfy (4.1a), we need to find factors $\Lambda_k^\delta \ge 0$ and penalties $\varepsilon_{k+1}^\delta \in \mathbb{R}$ such that

$$\frac{1}{2}\|x_\delta^k \circ v_\delta^k - \bar{x}^k \circ \bar{v}^k\|_X^2 \le \frac{\Lambda_k^\delta}{2}\|x_\delta^k - \bar{x}^k\|_X^2 + \varepsilon_{k+1}^\delta \quad (\bar{x}^k \in \mathcal{B}_k). \tag{5.3}$$

The satisfaction of (4.1b) is handled analogously. If we had no displacement field measurement error, i.e., $v_\delta^k = \bar{v}^k$, we could by the area formula take $\Lambda_k^\delta = \max_{\xi\in\Omega} |\det \nabla (v_\delta^k)^{-1}(\xi)|$ and $\varepsilon_{k+1}^\delta = 0$. Otherwise we need the more elaborate estimate of the next lemma.

Lemma 5.1. *Let $\bar{v} \in H^1(\mathbb{R}^2;\mathbb{R}^2)$ and $\bar{x} \in H^1(\Omega)$ with $\|\nabla \bar{x}\|_{2,\infty}^2 \le M$ for some $M > 0$. Let $\mathcal{V} \subset H^1(\Omega;\Omega)$ be a set of bijective displacement fields satisfying*

$$\Lambda_{\mathcal{V}} := \sup_{v\in\mathcal{V},\, \xi\in\Omega} |\det \nabla v^{-1}(\xi)| < \infty. \tag{5.4}$$

Then for any $x \in L^2(\Omega)$, $v \in \mathcal{V}$, and $\Lambda > \Lambda_{\mathcal{V}}$,

$$\frac{1}{2}\|x \circ v - \bar{x} \circ \bar{v}\|_{L^2(\Omega)}^2 \le \frac{\Lambda}{2}\|x - \bar{x}\|_{L^2(\Omega)}^2 + \frac{\Lambda_{\mathcal{V}}(4\Lambda - 3\Lambda_{\mathcal{V}})}{8(\Lambda - \Lambda_{\mathcal{V}})} M \|v - \bar{v}\|_{L^2(\Omega;\mathbb{R}^2)}^2.$$

Proof. By the area formula and Young's inequality, for any $t > 0$,

$$\begin{aligned}\int_\Omega |x(v) - \bar{x}(\bar{v})|^2 \, d\xi &\leq \int_\Omega \left(1 + \tfrac{t}{2}\right) |x(v(\xi)) - \bar{x}(v(\xi))|^2 + \left(1 + \tfrac{1}{2t}\right) |\bar{x}(v(\xi)) - \bar{x}(\bar{v}(\xi))|^2 \, d\xi \\ &= \left(1 + \tfrac{t}{2}\right) \int_\Omega |x(\xi) - \bar{x}(\xi)|^2 |\det \nabla v^{-1}(\xi)| \, d\xi + \left(1 + \tfrac{1}{2t}\right) \int_\Omega |\bar{x}(v) - \bar{x}(\bar{v})|^2 \, d\xi.\end{aligned}$$

Using (5.4) and that $\bar{x}$ is $\sqrt{M}$-Lipschitz, it follows

$$\|x \circ v - \bar{x} \circ \bar{v}\|^2_{L^2(\Omega)} \leq \left(1 + \tfrac{t}{2}\right) \Lambda_{\mathcal{V}} \|x - \bar{x}\|^2_{L^2(\Omega)} + \left(1 + \tfrac{1}{2t}\right) M \|v - \bar{v}\|^2_{L^2(\Omega;\mathbb{R}^2)}.$$

Taking $t = 2(\Lambda/\Lambda_{\mathcal{V}} - 1)$ yields the claim. □

We need the primal iterates to stay bounded. For this we use the next lemma:

Lemma 5.2. *Compute x_δ^k and v_δ^k by Algorithm 4.1 for (5.1a) with fixed $\delta > 0$ and $\tau_k \equiv \tau > 0$. Suppose $\tau \leq \frac{(2-\Lambda)C-\varepsilon}{\alpha^2 \|D\|^2}$ and $\|\xi_\delta^k - b_\delta^k\|^2 \leq C\Lambda + \varepsilon$ for some $C, \Lambda, \varepsilon > 0$. Then $\|x_\delta^k - b_\delta^k\|^2 \leq C$.*

Proof. We drop the indexing by δ as it is fixed. The dual prediction of Algorithm 4.1 guarantees $\|v^k\|_{2,\infty} \leq \alpha$. The primal step is

$$x^k := \arg\min_x \|x - \xi^k - \tau D v^k\|^2 + \tau \|x - b^k\|^2. \tag{5.5}$$

The optimality conditions are $0 = x^k - \xi^k + \tau D v^k + \tau(x^k - b^k)$. Thus $\tau \|x^k - b^k\| = \|x^k - \xi^k + \tau D v^k\|$. By (5.5), comparing to $x = \xi^k$, we get

$$2\|x^k - b^k\|^2 \leq \tau \|D v^k\|^2 + \|\xi^k - b^k\|^2 \leq \tau \alpha^2 \|D\|^2 + C\Lambda + \varepsilon.$$

Thus $\|x^k - b^k\|^2 \leq C$ when τ is as stated. □

We may now prove convergence to the true data as the displacement field measurement error $\varepsilon \searrow 0$ along with the noise in the data b_δ^k.

Theorem 5.3. *For all $k \in \mathbb{N}$, $\delta > 0$, and some $\alpha = \alpha(\delta) \searrow 0$ as $\delta \searrow 0$, assume the setup of (5.1) and (5.2) with $v_\delta^k \in \mathcal{V}$ for a set $\mathcal{V} \subset V$ of bijective displacement fields such that $\Lambda_{\mathcal{V}} < 2$. With $\bar{b}^{0:\infty} \in \mathcal{B}$, assume:*

- *(I) $\sup_{k \in \mathbb{N}} \|b_\delta^k - \bar{b}^k\|_{L^2(\Omega)} \to 0$ and $\sup_{k \in \mathbb{N}} \|v_\delta^k - \bar{v}^k\|_{L^2(\Omega;\mathbb{R}^2)} \to 0$ as $\delta \searrow 0$.*
- *(II) For some $\Lambda_k, \Theta_k \equiv \Lambda > \Lambda_{\mathcal{V}}$, the step length parameters are as in Example 4.10, independent of δ and k.*

For an initial $u^0 = u_\delta^0$, for all $\delta > 0$, generate $u_\delta^{1:\infty}$ by Algorithm 4.1. Then there exist $\bar{N}(\delta) \in \mathbb{N}$ such that:

- *(a) $\lim_{\delta \searrow 0} \sup_{N \geq \bar{N}(\delta)} \frac{1}{N} \sum_{k=1}^N \|x_\delta^k - \bar{b}^k\|^2_{L^2(\Omega)} = 0$, and*
- *(b) provided $\tau\sigma\|D\|^2 < 1$, moreover, $\lim_{\delta \searrow 0} \sup_{N \geq \bar{N}(\delta)} \frac{1}{2N} \sum_{k=0}^{N-1} \|x_\delta^{k+1} - x_\delta^k \circ v_\delta^k\|^2_{L^2(\Omega)} = 0$.*

Proof. We first show the boundedness of $\{x_\delta^k\}_{k \in \mathbb{N}, \delta \in (0,\bar{\delta})}$ for some $\bar{\delta} > 0$. By (I), $\sup_{k \in \mathbb{N}} \|b_\delta^k - \bar{b}^k\|_X =: \delta_b \to 0$ and $\sup_{k \in \mathbb{N}} \|v_\delta^k - \bar{v}^k\|_{L^2(\Omega;\mathbb{R}^2)} =: \delta_v \to 0$ as $\delta \searrow 0$. We have $\bar{b}^{k+1} = \bar{b}^k \circ \bar{v}^k$ and $\xi_\delta^{k+1} = x_\delta^k \circ v_\delta^k$. Using Young's inequality twice for any $\beta > 0$ and Lemma 5.1 for any $\Lambda' > \Lambda_{\mathcal{V}}$,

$$\begin{aligned}\|\xi_\delta^{k+1} - b_\delta^{k+1}\|^2_{L^2(\Omega)} &\leq (1+\beta)\|x_\delta^k \circ v_\delta^k - \bar{b}^k \circ \bar{v}^k\|^2_{L^2(\Omega)} + (1+\beta^{-1})\|b_\delta^{k+1} - \bar{b}^k \circ \bar{v}^k\|^2_{L^2(\Omega)} \\ &\leq \Lambda'(1+\beta)\|x_\delta^k - \bar{b}^k\|^2_{L^2(\Omega)} + (1+\beta^{-1})\delta_b^2 + (1+\beta)\tfrac{\Lambda_{\mathcal{V}}(4\Lambda' - 3\Lambda_{\mathcal{V}})}{8(\Lambda' - \Lambda_{\mathcal{V}})} M \delta_v^2 \\ &\leq \Lambda'(1+\beta)^2 \|x_\delta^k - b_\delta^k\|^2_{L^2(\Omega)} + (1+\beta^{-1})\left(\Lambda'(1+\beta) + 1\right)\delta_b^2 + (1+\beta)\tfrac{\Lambda_{\mathcal{V}}(4\Lambda' - 3\Lambda_{\mathcal{V}})}{8(\Lambda' - \Lambda_{\mathcal{V}})} M \delta_v^2.\end{aligned}$$

Taking $\beta > 0$, $\Lambda' > \Lambda_{\mathcal{V}}$ small enough, we obtain for any $\Lambda \in (\Lambda_{\mathcal{V}}, 2)$ that $\|\xi_\delta^{k+1} - b_\delta^{k+1}\|^2 \leq \Lambda \|x_\delta^k - b_\delta^k\|^2 + \varepsilon_\delta$ for some $\varepsilon_\delta \searrow 0$ as $\delta \searrow 0$. By Lemma 5.2, now $\sup_k \|x_\delta^k - b_\delta^k\|^2 \leq C$ for any $C \geq \|x^0 - b_\delta^0\|^2$ with $\tau \leq \frac{(2-\Lambda)C - \varepsilon_\delta}{\alpha^2 \|D\|^2}$. This holds for C large and $\delta \in (0, \bar\delta)$ for small $\bar\delta > 0$. Thus $\sup_{k\in\mathbb{N}, \delta \in (0,\bar\delta)} \|x_\delta^k\| < \infty$.

Fix now $\delta \in (0, \bar\delta)$ and $N \geq 1$. By Lemma 5.1 and (II), the prediction bounds (4.1) hold for all $k \in \mathbb{N}$ with $\Lambda_k = \Theta_k \equiv \Lambda$ and

$$\varepsilon_{k+1} = \tilde\varepsilon_{k+1} = \varepsilon_{k+1}^\delta := \frac{\Lambda_{\mathcal{V}}(4\Lambda_k - 3\Lambda_{\mathcal{V}})}{8(\Lambda_k - \Lambda_{\mathcal{V}})} M \|v_\delta^k - \bar v^k\|_{L^2(\Omega;\mathbb{R}^2)}^2 \leq \frac{\Lambda_{\mathcal{V}}(4\Lambda - 3\Lambda_{\mathcal{V}})}{8(\Lambda - \Lambda_{\mathcal{V}})} M \delta_v. \tag{5.6}$$

The rest of Assumption 4.1 holds by the construction in (5.1) and (5.2) while (4.5) holds by (II) and Example 4.10. By (4.8) and Example 4.10, also $Z_k M_k \geq \left(\begin{smallmatrix} 1-\tau\sigma\|D\|^2 & 0 \\ 0 & 0 \end{smallmatrix}\right) \geq 0$. Therefore, by Theorem 4.6, for some constant $C > 0$ (dependent on the initialisation, $\sup_{\delta\in(0,\bar\delta)} \delta_v$, $\Theta_k \equiv \Lambda$, and $\rho_k \equiv 0$ as well as $\varphi_k \equiv 1$ and $\psi_k \equiv \frac{\tau}{\sigma}$ as in Example 4.10), we have with the notation $F_{1:N}$ etc. from Theorem 4.6 that

$$\frac{1}{N}[F_{1:N}^\delta + \check G_{1:N}^{\alpha(\delta)} \circ K_{1:N}](x_\delta^{1:N}) - \frac{1}{N} \inf_{\bar x^{1:N} \in \mathcal{B}_{1:N}} [F_{1:N}^\delta + G_{1:N}^{\alpha(\delta)} \circ K_{1:N}](\bar x^{1:N}) + \sum_{k=0}^{N-1} \frac{1 - \tau\sigma\|D\|}{2N} \|x_\delta^{k+1} - x_\delta^k \circ v_\delta^k\|^2 \leq \frac{C}{N}.$$

By Lemma 3.7, the defining (5.1), and the just proved boundedness of the iterates,

$$\check G_{1:N}^{\alpha(\delta)}(K_{1:N} x_\delta^{1:N}) \geq -G_{1:N}^{\alpha(\delta)}(-K_{1:N} x_\delta^{1:N}) = -\sum_{k=1}^N \alpha(\delta) \|D x_\delta^k\|_B \geq -\alpha(\delta) N C'$$

for some constant $C' > 0$. Since $\mathcal{B}$ is bounded (by the boundedness of $\mathcal{U}_0$ and finite-dimensionality), also $[G_{1:N}^{\alpha(\delta)} \circ K_{1:N}](\bar x^{1:N}) \leq \alpha \bar C'$ for some $\bar C' > 0$. Hence, for some $C'' > 0$ we get for all $\bar x^{1:N} \in \mathcal{B}_{1:N}$ that

$$\sum_{k=1}^N \left(\tau \frac{F_k^\delta(x_\delta^k) - F_k^\delta(\bar x^k)}{N} + \frac{1 - \tau\sigma\|D\|}{2N} \|x_\delta^{k+1} - x_\delta^k \circ v_\delta^k\|^2 \right) \leq \alpha(\delta) C'' + \frac{C}{N}.$$

Due to (5.2b), this says for all $\delta \in (0, \bar\delta)$ and $N \geq 1$ that

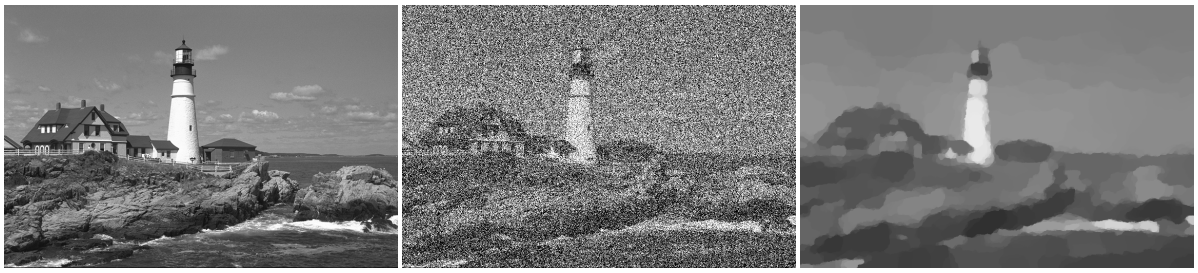
$$\sum_{k=1}^N \left(\frac{\tau}{2N} \|b_\delta^k - x^k\|^2 + \frac{1 - \tau\sigma\|D\|}{2N} \|x_\delta^{k+1} - x_\delta^k \circ v_\delta^k\|^2 \right) \leq \sum_{k=1}^N \frac{\tau}{2N} \|b_\delta^k - \bar b^k\|^2 + \alpha(\delta) C'' + \frac{C}{N}.$$

Since $\alpha(\delta) \searrow 0$ and (I) guarantees $\sup_{k\in\mathbb{N}} \|b_\delta^k - \bar b^k\|_{L^2(\Omega)} \to 0$ as $\delta \searrow 0$, the right hand side can be made smaller than δ by taking $N \geq N(\delta)$ large enough. The claim (b) immediately follows while (a) follows after further referral to (I). □

NUMERICAL SETUP

We perform our experiments on a simple square image as well as the lighthouse image from the free Kodak image suite [16]; this is in Figure 1 along with a noisy version and comparison single-frame total variation reconstruction. The original size is 768×512 pixels. For our experiments, we pick a 300×200 subimage moving according to Brownian motion of standard deviation 2. Thus the displacement fields $\bar v^k(\xi) = \xi - \bar u^k$ with $\bar u^k \in \mathbb{R}^2$ are constant in space. To the subimage we add 50% Gaussian noise (standard deviation 0.5 with original intensities in $[0, 1]$). To construct the measured displacements available to the algorithm we add 5% Gaussian noise (standard deviation $0.05\|\bar u^k\|$) to the true displacements.[4]

[4] Then (5.4) gives $\Lambda_{\mathcal{V}} = 1$. Constant true displacements are allowed by Lemma 5.1, but constant measurements not. If $\|x - \bar x\|_{L^2(\Omega + B(0,\|u\|))}^2 \leq C\|x - \bar x\|_{L^2(\Omega)}^2$ then Lemma 5.1 and Theorem 5.3 extend to $\Lambda > C\Lambda_{\mathcal{V}}$. In practise, to compute $x \circ v$, we extrapolate x outside Ω such that Neumann boundary conditions are satisfied.



(a) Original (b) Data: noise level 0.5 (c) Recon.: noise level 0.5, $\alpha = 1$

Figure 1: Test image, added noise, and stationary reconstruction for comparison.

We take the regularisation parameter $\alpha = 1$. The corresponding full-image total variation reconstruction is in Figure 1c. To parametrise the POPD (Algorithm 4.1) we

- Fix the primal step length parameter $\tau = 0.01$ as well as $\Lambda = \Theta = 1$ and $\kappa = 0.9$.
- Take the primal strong convexity factor $\gamma = 1$ and generally the dual factor $\rho = 0$.
- Take $\tilde{\sigma}$, maximal σ, and minimal $\tilde{\rho}_{k+1} \equiv \tilde{\rho}$ according to Example 4.10. Here we estimate $\|K_k\| \leq \sqrt{8}$ for forward-differences discretisation of $K_k = D$ with cell width $h = 1$ [10].

Although G^*_{k+1} is not strongly convex, we also experiment taking a "phantom" $\rho = 100$. This can in principle be justified via *local* strong convexity or *strong metric subregularity* at a solution. We briefly indicate how this works in Appendix A. The effect in practise is to increase the dual step length parameter σ. We always take zero as the initial iterate (primal and dual).

We implemented our algorithms in Julia 1.3 [7], and performed our experiments on a mid-2014 MacBook Pro with 16GB RAM and two CPU cores. Our implementation uses a maximum of four computational threads (two cores with hyperthreading) in those parts of the code where this appears advantageous. The data generation runs in its own thread. The implementation is available on Zenodo [32].

NUMERICAL RESULTS

We display the reconstructions in Figures 2 to 4 and the performance (function value, PSNR, and SSIM) in Figures 6a, 7a and 8a. The reconstructions are for the frames/iterations 30, 50, 100, 300, 500, 1000, and 3000 whereas the performance plots display all 10000 iterations at a resolution of 100 iterations after the first 100 iterations. The right-most column of the reconstruction figures displays the true cumulative displacement field up to the corresponding data frame (indicated in the bottom-left corner). The darker line is sampled at the same resolution as the performance plots whereas the lighter line is sampled at every iteration. Regarding real-time computability, averaged over the 10000 iterations, every iteration takes ~6.5ms, which is to say the POPD can process 154 frames per second.

The performance plots show convergence of the function value to a stable value, not necessarily a minimum, within 100 iterations. Likewise the SSIM and PSNR reach a relatively stable and acceptable value by 100 iterations. Visually, we have decent tracking of movement, but we need the large ρ-value to get a noticeable cartoon-like "total variation effect". In the last frame of Figure 2 we can see the effect of the algorithm not being able to track a sudden large displacement fast enough, hence producing some motion blur. The 100 iterations, that were needed to reach a stable function value, SSIM, or PSNR, appear to be mainly needed to reach the correct contrast level: recall that we initialise with zero. We tested initialising the primal variable with the noisy data: the algorithm then needed a similar number of iterations to reduce the noise. A smarter initialisation might help reduce the 100-iteration "initialisation window".

For comparison, we have included POFB reconstruction (Algorithm 2.1) in Figure 5. We use the step

length parameter $\tau = 0.01$ for the POFB itself. We take 10 iterations of FISTA [4] with step length parameter $\tilde{\tau} = 1/\|K\|^2$ to approximately solve the proximal step. By the performance measures the results are comparable to the POPD. Visually they are similar to the high-ρ POPD. The algorithm is, however, quite a bit slower: ~21.2ms/frame or 47 frames per second. Solving the proximal step accurately would further slow it down.

5.2 UNKNOWN DISPLACEMENT FIELD

When the displacement field v_k is completely unknown, we need to estimate it from data. For some $E_k : V \to \overline{\mathbb{R}}$ we do this through

$$\min_{x\in X,\, v\in V} \frac{1}{2}\|b_k - x\|_X^2 + \alpha\|Dx\| + E_k(v) \tag{5.7}$$

We drop the indexing by the noise level $\delta > 0$ as we will not be studying regularisation properties. Ideally we would take $E_k(v)$ as $\frac{\theta}{2}\|b^{k+1} - b^k \circ v\|_X^2$, plus regularisation terms. However, the resulting problem would be highly nonconvex. A second idea is to use a Horn–Schunck [20] type penalty on linearised optical flow[5], taking for some parameters $\theta, \lambda_1, \lambda_2 > 0$,

$$E_k(v) = \frac{\theta}{2}\|b_{k+1} - b_k + \langle\!\langle \mathrm{Id} - v, \nabla b_k\rangle\!\rangle\|_X^2 + \frac{\lambda_1}{2}\|\,\mathrm{Id} - v\|_2^2 + \frac{\lambda_2}{2}\|\nabla v\|_2^2, \tag{5.8}$$

where the pointwise inner product $\langle\!\langle a, b\rangle\!\rangle(\xi) := \langle a(\xi), b(\xi)\rangle$. We regularise the displacement field v to both be close to identity (no displacement) and to be smooth in space.[6]

The choice (5.8) is, however, very inaccurate in practise. We therefore, firstly, introduce a time-step parameter T and a convolution kernel ϱ to counteract noise in the data. Secondly, we average the Horn–Schunck term over a window of n frames. For iteration k, the last frame is

$$\iota(k) := \max\{1, k + 1 - (n-1)\} \quad \text{and its true length} \quad n_k := k + 1 - (\iota(k) - 1).$$

With $j \in \{\iota(k), \dots, k+1\}$, we write $v_j^k \in V$ for the displacement of b^j from $b^{\iota(k)-1}$ as estimated on iteration k. Then the displacement of b^{j+1} from b^j is $(v_j^k)^{-1} \circ v_{j+1}^k$. We take $E_k : V^{n_{k+1}} \to \mathbb{R}$,

$$\begin{aligned} E_k(v^k_{\iota(k):k+1}) := \frac{1}{n_k}\sum_{j=\iota(k)-1}^{k} \Big(&\frac{\theta}{2}\|\varrho * (b^{j+1} - b^j)/T + \langle\!\langle \mathrm{Id} - (v_j^k)^{-1} \circ v_{j+1}^k, \nabla(\varrho * b^j)\rangle\!\rangle\|_X^2 \\ &+ \frac{\lambda_1}{2}\|\,\mathrm{Id} - (v_j^k)^{-1} \circ v_{j+1}^k\|_2^2 + \frac{\lambda_2}{2}\|\nabla v_j^k\|_2^2\Big). \end{aligned} \tag{5.9}$$

Although not given as a parameter, we use $v^k_{\iota(k)-1} = 0$.

We predict the primal variables using

$$A_k(x, v^k_{\iota(k):k+1}) := \begin{cases} (x \circ (v_k^k)^{-1} \circ v_{k+1}^k, v_1^k, \dots, v_{k+1}^k, 0), & k < n, \\ (x \circ (v_k^k)^{-1} \circ v_{k+1}^k, (v^k_{\iota(k)})^{-1} \circ v^k_{\iota(k+1)}, \dots, (v^k_{\iota(k)})^{-1} \circ v^k_{k+1}, 0), & k \geq n, \end{cases}$$

and the dual variables using

$$B_k(y) := y \circ (v_k^k)^{-1} \circ v_{k+1}^k$$

[5] To obtain the linearised optical flow model, we start with $b_{k+1}(\xi) = b_k(v_k(\xi))$ holding for all $\xi \in \Omega$ and a sufficiently smooth image b_k. By Taylor expansion $b_k(v_k(\xi)) \approx b_k(\xi) + \langle \nabla b_k(\xi), v_k(\xi) - \xi\rangle$. Thus $0 = b_{k+1}(\xi) - b_k(v_k(\xi)) \approx b_{k+1}(\xi) - b_k(\xi) + \langle \nabla b_k(\xi), \xi - v_k(\xi)\rangle$.

[6] Indeed, in linearised optical flow the displacement field cannot in general be discontinuous. See [29, 13] for approaches designed to avoid this restriction.

Hence we a) propagate the image x and the dual variable using the estimated displacement of the next frame from the current frame, b) update the displacement estimates to be with respect to the start $\iota(k+1)$ of the new n-frame window, and c) predict the displacement between the next two frames to be zero. The latter is consistent with the zero-mean Brownian motion used in our numerical experiments.

We write the problem (5.7) with E_k given by (5.9) in the form (1.6) by taking

$$F_k(x, v^k_{\iota(k):k+1}) := \frac{1}{2}\|b^k - x\|^2_V + E_k(v^k_{\iota(k):k+1}), \quad K_k(x, v^k_{\iota(k):k+1}) := Dx, \quad \text{and} \quad G^*_k(y) := \delta_{\alpha B}(y).$$

We split $\operatorname{prox}_{\tau F_k}$ into individual updates with respect to x and $v^k_{\iota(k):k+1}$. If the displacement fields are constant in space, $v^k_j(\xi) = \xi - u^k_j$ with $u^k_j \in \mathbb{R}^2$, the compositions $(v^k_{\iota(k)})^{-1} \circ v^k_j \equiv u^k_{\iota(k)} - u^k_j$, and $\operatorname{prox}_{\tau E_k}$ reduces to an easily solvable chain of 2×2 quadratic optimisation problems.

The Horn–Schunck linearisation of the optical flow only converges to the true optical flow as we increase the temporal resolution. Therefore, an equivalent of the regularisation theory of Theorem 5.3 for the present model would require increasing the temporal resolution as $\delta \searrow 0$ and $N \nearrow \infty$. As the analysis is somewhat involved, we have decided not to pursue such estimates. It is, however, not difficult to extend the prediction bounds of Lemma 5.1.

NUMERICAL SETUP AND RESULTS

For our numerical experiments we use generally the same setup as in Section 5.1 except we reduce the noise level in the image to 30% and correspondingly take $\alpha = 0.2$. For our new parameters we take $\lambda_1 = 1$ and $\theta = (300 \cdot 200) \cdot 100^3$ with constant-in-space displacement fields, so that λ_2 is irrelevant in (5.9). For the displacement estimation we use a window of $n = 100$ previous frames. For the smoothing kernel ϱ in the Horn–Schunck term of (5.9) we take a normalised Gaussian of standard deviation 3 pixels in a window of 11×11 pixels. We also take the time step parameter $T = 0.5$ for the lighthouse and $T = 1$ for the square test image. Our Julia implementation is available on Zenodo [32].

The reconstructions and estimated displacements are in Figures 9 to 11 and the performance plots (function value, PSNR, SSIM) in Figures 6b, 7b and 8b. Regarding real-time computability, the POPD requires 20.8ms/iteration, that is, can process 48 frames per second.

The function values take a long time to decrease. The PSNR and SSIM, however, again reach an acceptable and somewhat stable value after 100–200 iterations. Visually, the results are somewhat more blurred than with the approximately known displacement in Section 5.1, and even with $\rho = 100$ the cartoon-like total variation effect remains small. Nevertheless, the reconstructions are visually pleasing and the displacement is estimated to an acceptable accuracy. This did, however, require adapting the time-step parameter T to the test case. Improving the optical flow model to not require such an extraneous parameter is something for future research: we believe that the present results already demonstrate that online optimisation is a worthy approach to dynamic imaging.

6 CONCLUSION

With the goal of solving—for now relatively simple—imaging problems "online", in real-time, we incorporated predictors into the forward-backward and primal-dual proximal splitting methods. For the predictive online forward-backward method (POFB) a reasonable notion of "dynamic regret" stays bounded, and can even converge below zero. Using regularisation theory we, moreover, proved convergence to a ground-truth as the level of corruption in the problem data vanishes. Hence the method forms an appropriate regulariser.

We do not, yet, understand the predictive online primal-dual method (POPD) as well. While we have shown analogous results, including convergence as the data improves, the form of "regret" we were able to employ still requires study and interpretation. This notwithstanding, our numerical results on

optical flow are encouraging. More research is needed to understand the parametrisation and improved predictors needed to make the total variation effect prominent.

APPENDIX A LOCAL STRONG CONVEXITY

We establish local strong convexity of the indicator function of the ball. This has been shown in [1] to be equivalent to the *strong metric subregularity* of the subdifferential. For related characterisations, see also [33] and regarding total variation [22, appendix].

Lemma A.1. *With $F : X \to \mathbb{R}, F = \delta_{\operatorname{cl} B(0,\alpha)}$ on a Hilbert space X, suppose $x \in \partial B(0,\alpha)$ and $0 \neq x^* \in \partial F(x)$. Then*

$$F(x') - F(x) \geq \langle x^*, x' - x \rangle + \frac{\gamma}{2}\|x' - x\|^2 \quad (x' \in U_x)$$

for

$$U_x = \begin{cases} X, & 0 \leq \gamma\alpha \leq \|x^*\|, \\ [\operatorname{cl} B(0,\alpha)]^c \cup \operatorname{cl} B(x,\alpha), & \alpha\gamma > \|x^*\|. \end{cases}$$

Proof. Observe that $x^* = \lambda x$ for $\lambda := \|x^*\|/\alpha$. If $x' \notin \operatorname{cl} B(0,\alpha)$, there is nothing to prove. So take $x' \in \operatorname{cl} B(0,\alpha)$. Then we need $0 \geq \lambda\langle x, x' - x\rangle + \frac{\gamma}{2}\|x' - x\|^2$. Since $\|x\| = \alpha$, this says

$$\left(\lambda - \frac{\gamma}{2}\right)\alpha^2 \geq \frac{\gamma}{2}\|x'\|^2 + (\lambda - \gamma)\,\langle x, x'\rangle. \tag{A.1}$$

Suppose $\gamma \leq \lambda$, which is the first case of U_x. Then (A.1) is seen to hold by application of Young's inequality on the inner product term, followed by $\|x'\| \leq \alpha$.

If on the other hand, $\gamma > \lambda$, which is the second case of U_x, we take $x' \in \operatorname{cl} B(x,\alpha) \cap \operatorname{cl} B(0,\alpha)$. This implies $\langle x', x\rangle \geq \frac{1}{2}\|x'\|^2$. Since $\lambda - \gamma < 0$, this and $\|x'\| \leq \alpha$ prove (A.1). □

REFERENCES

[1] F. J. Aragón Artacho and M. H. Geoffroy, Characterization of metric regularity of subdifferentials, *Journal of Convex Analysis* 15 (2008), 365–380.

[2] N. Bastianello, A. Simonetto, and R. Carli, Prediction-Correction Splittings for Time-Varying Optimization With Intermittent Observations, *IEEE Control Systems Letters* 4 (2020), 373–378, doi:10.1109/lcsys.2019.2930491.

[3] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer, 2 edition, 2017, doi:10.1007/978-3-319-48311-5.

[4] A. Beck and M. Teboulle, A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems, *SIAM Journal on Imaging Sciences* 2 (2009), 183–202, doi:10.1137/080716542.

[5] F. Becker, S. Petra, and C. Schnörr, Optical Flow, in *Handbook of Mathematical Methods in Imaging*, O. Scherzer (ed.), Springer, 2015, 1945–2004, doi:10.1007/978-1-4939-0790-8_38.

[6] E. V. Belmega, P. Mertikopoulos, R. Negrel, and L. Sanguinetti, Online convex optimization and no-regret learning: Algorithms, guarantees and applications, 2018, arXiv:804.04529.

[7] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, Julia: A fresh approach to numerical computing, *SIAM Review* 59 (2017), 65–98, https://doi.org/10.1137/141000671.

[8] L. Biegler, O. Ghattas, M. Heinkenschloss, D. Keyes, and B. Waanders, *Real-Time PDE-Constrained Optimization*, Computational Science and Engineering, SIAM, 2007.

[9] O. Bousquet and L. Bottou, The Tradeoffs of Large Scale Learning, *Advances in Neural Information Processing Systems* 20 (2008), 161–168, http://papers.nips.cc/paper/3323-the-tradeoffs-of-large-scale-learning.pdf.

[10] A. Chambolle, An algorithm for total variation minimization and applications, *Journal of Mathematical Imaging and Vision* 20 (2004), 89–97, doi:10.1023/b:jmiv.0000011325.36760.1e.

[11] A. Chambolle and T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging, *Journal of Mathematical Imaging and Vision* 40 (2011), 120–145, doi:10.1007/s10851-010-0251-1.

[12] K. Chaudhury and R. Mehrotra, A trajectory-based computational model for optical flow estimation, *IEEE Transactions on Robotics and Automation* 11 (1995), 733–741, doi:10.1109/70.466611.

[13] K. Chen and D. A. Lorenz, Image Sequence Interpolation Based on Optical Flow, Segmentation, and Optimal Control, *IEEE Transactions on Image Processing* 21 (2012), doi:10.1109/tip.2011.2179305.

[14] C. Clason and T. Valkonen, Introduction to Nonsmooth Analysis and Optimization, 2020, arXiv:2001.00216, https://tuomov.iki.fi/m/nonsmoothbook_part.pdf. Work in progress.

[15] H. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Mathematics and Its Applications, Springer, 2000.

[16] R. Franzen, Kodak lossless true color image suite, PhotoCD PCD0992. Lossless, true color images released by the Eastman Kodak Company, 1999, http://rok.us/graphics/kodak/.

[17] M. Grötschel, S. Krumke, and J. Rambau, *Online Optimization of Large Scale Systems*, Springer, 2013.

[18] E. Hall and R. Willett, Dynamical models and tracking regret in online convex programming, in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta and D. McAllester (eds.), volume 28 of Proceedings of Machine Learning Research, PMLR, Atlanta, Georgia, USA, 2013, 579–587, http://proceedings.mlr.press/v28/hall13.html.

[19] E. Hazan, Introduction to Online Convex Optimization, *Foundations and Trends in Optimization* 2 (2016), 157–325, doi:10.1561/2400000013.

[20] B. K. Horn and B. G. Schunck, Determining Optical Flow, in *Proc. SPIE*, volume 0281, SPIE, 1981, 319–331, doi:10.1117/12.965761.

[21] J. A. Iglesias and C. Kirisits, Convective regularization for optical flow, in *Variational Methods In Imaging and Geometric Control*, De Gruyter, 2016, 184–201, doi:10.1515/9783110430394.

[22] J. Jauhiainen, P. Kuusela, A. Seppänen, and T. Valkonen, Relaxed Gauss–Newton methods with applications to electrical impedance tomography, *SIAM Journal on Imaging Sciences* (2020), arXiv:2002.08044, https://tuomov.iki.fi/m/gn_overrelax.pdf. in press.

[23] H. H. Nagel, Extending the 'Oriented smoothness constraint' into the temporal domain and the estimation of derivatives of optical flow, in *Computer Vision—ECCV 90*, O. Faugeras (ed.), Springer, Berlin, Heidelberg, 1990, 139–148.

[24] H. H. Nagel et al., Constraints for the Estimation of Displacement Vector Fields From Image Sequences, in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (II)*, volume 2, IJCAI, 1983, 945–951.

[25] F. Orabona, A Modern Introduction to Online Learning, 2020, arXiv:1912.13213.

[26] A. Salgado and J. Sánchez, Temporal Constraints in Large Optical Flow Estimation, in *Computer Aided Systems Theory–EUROCAST 2007*, R. Moreno Díaz, F. Pichler, and A. Quesada Arencibia (eds.), Springer, Berlin, Heidelberg, 2007, 709–716.

[27] A. Simonetto, Time-varying convex optimization via time-varying averaged operators, 2017, arXiv:1704.07338.

[28] M. Tico, Digital Image Stabilization, in *Recent Advances in Signal Processing*, A. A. Zaher (ed.), IntechOpen, Rijeka, 2009, chapter 1, doi:10.5772/7458.

[29] T. Valkonen, Transport equation and image interpolation with SBD velocity fields, *Journal de mathématiques pures et appliquées* 95 (2011), 459–494, doi:10.1016/j.matpur.2010.10.010, https://tuomov.iki.fi/m/bd.pdf.

[30] T. Valkonen, Testing and non-linear preconditioning of the proximal point method, *Applied Mathematics and Optimization* (2018), doi:10.1007/s00245-018-9541-6, arXiv:1703.05705, https://tuomov.iki.fi/m/proxtest.pdf.

[31] T. Valkonen, First-order primal-dual methods for nonsmooth nonconvex optimisation, 2019, arXiv:1910.00115, https://tuomov.iki.fi/m/firstorder.pdf. submitted.

[32] T. Valkonen, Julia codes for "Predictive online optimisation with applications to optical flow", Software on Zenodo, 2020, doi:10.5281/zenodo.3659180.

[33] T. Valkonen, Preconditioned proximal point methods and notions of partial subregularity, *Journal of Convex Analysis* (2020), arXiv:1711.05123, https://tuomov.iki.fi/m/subreg.pdf. in press.

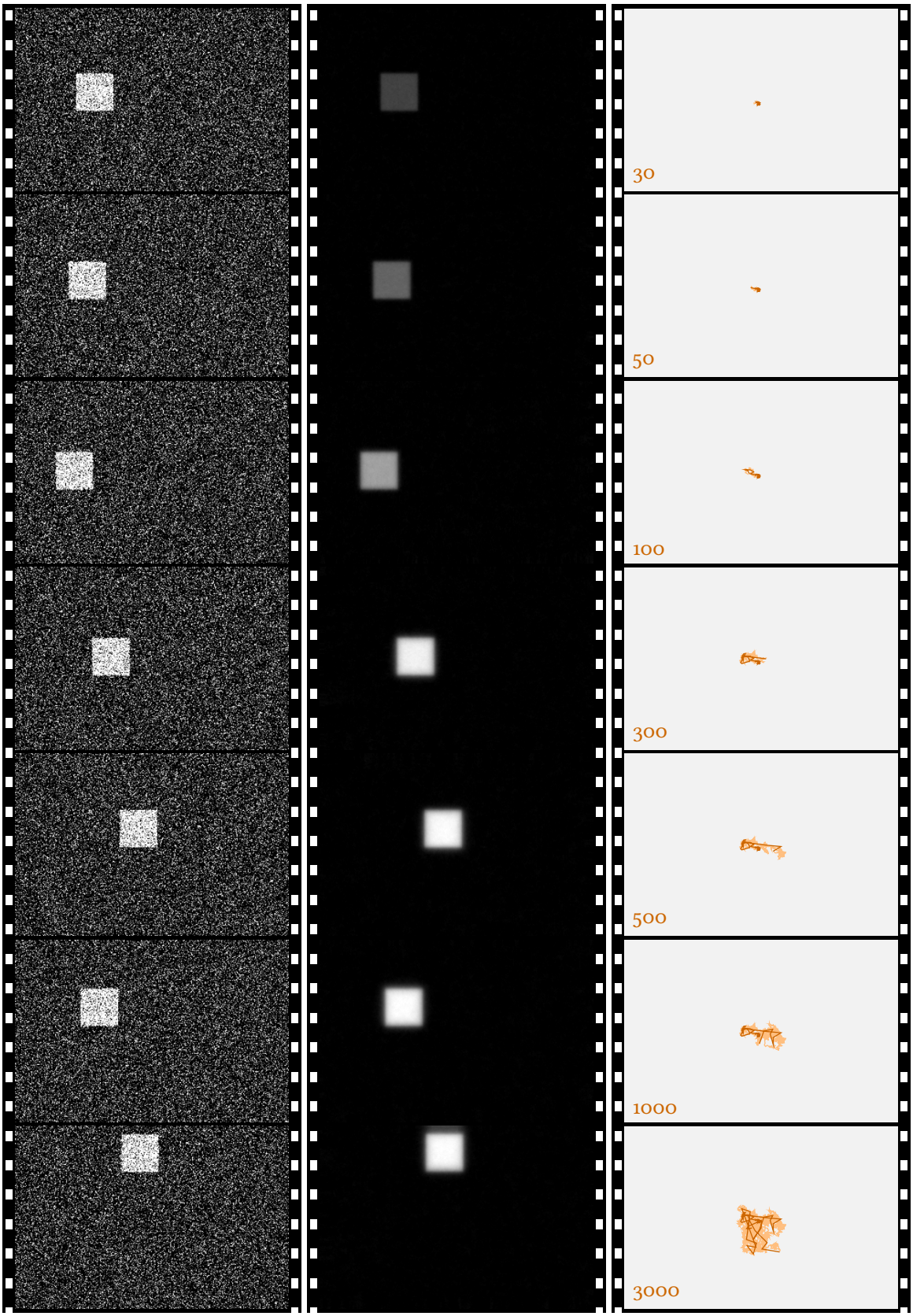
[34] S. Volz, A. Bruhn, L. Valgaerts, and H. Zimmer, Modeling temporal coherence for optical flow, in *2011 International Conference on Computer Vision*, IEEE, 2011, 1116–1123, doi:10.1109/iccv.2011.6126359.

[35] J. Weickert and C. Schnörr, Variational Optic Flow Computation with a Spatio-Temporal Smoothness Constraint, *Journal of Mathematical Imaging and Vision* 14 (2001), 245–255, doi:10.1023/a:1011286029287.

[36] Y. Zhang, R. J. Ravier, V. Tarokh, and M. M. Zavlanos, Distributed Online Convex Optimization with Improved Dynamic Regret, 2019, arXiv:1911.05127.

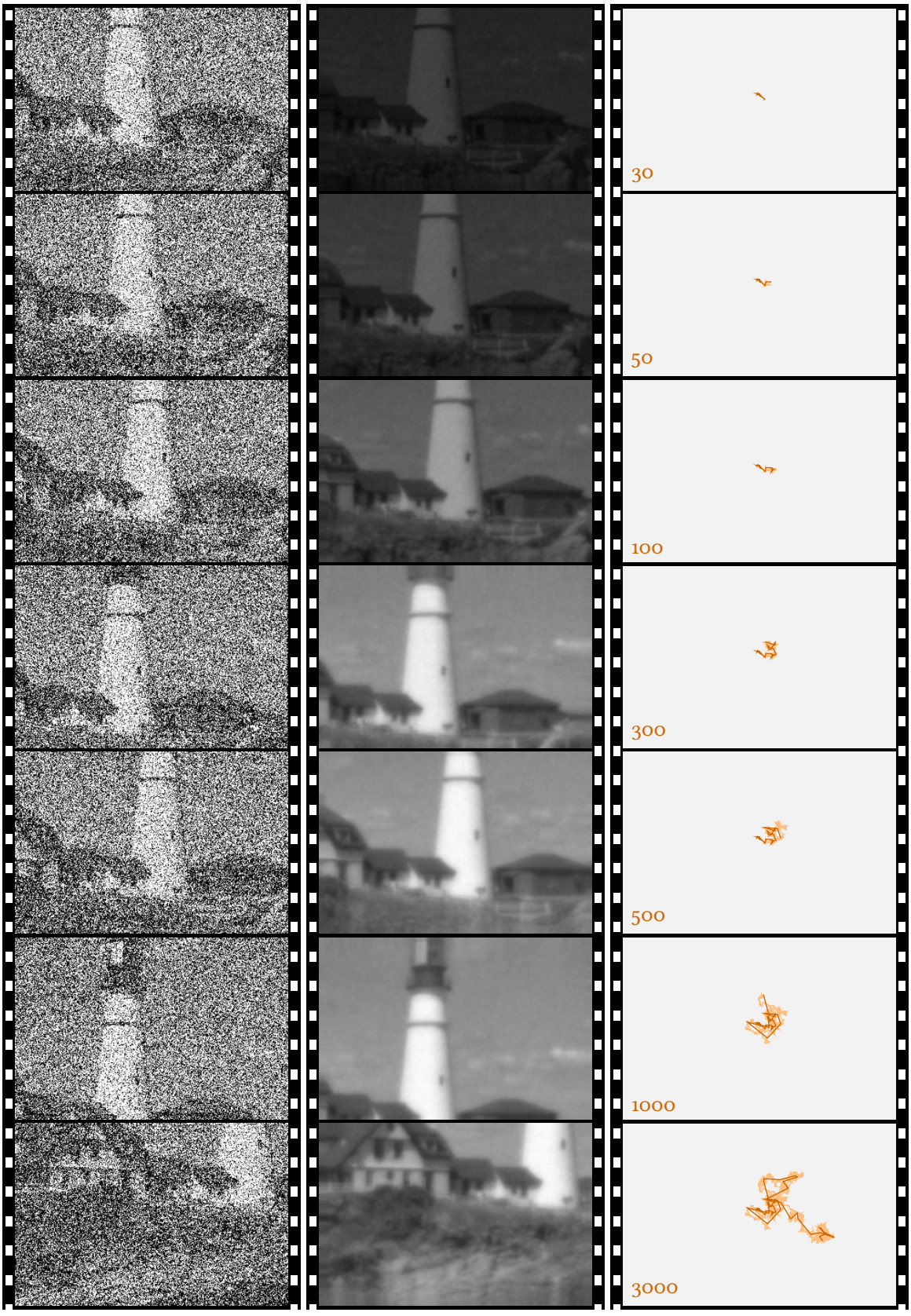
[37] J. Zhou, P. Hubel, M. Tico, A. N. Schulze, and R. Toft, Image registration methods for still image stabilization, US Patent 9,384,552, 2016.

[38] M. Zinkevich, Online convex programming and generalized infinitesimal gradient ascent, in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, AAAI, 2003, 928–936.



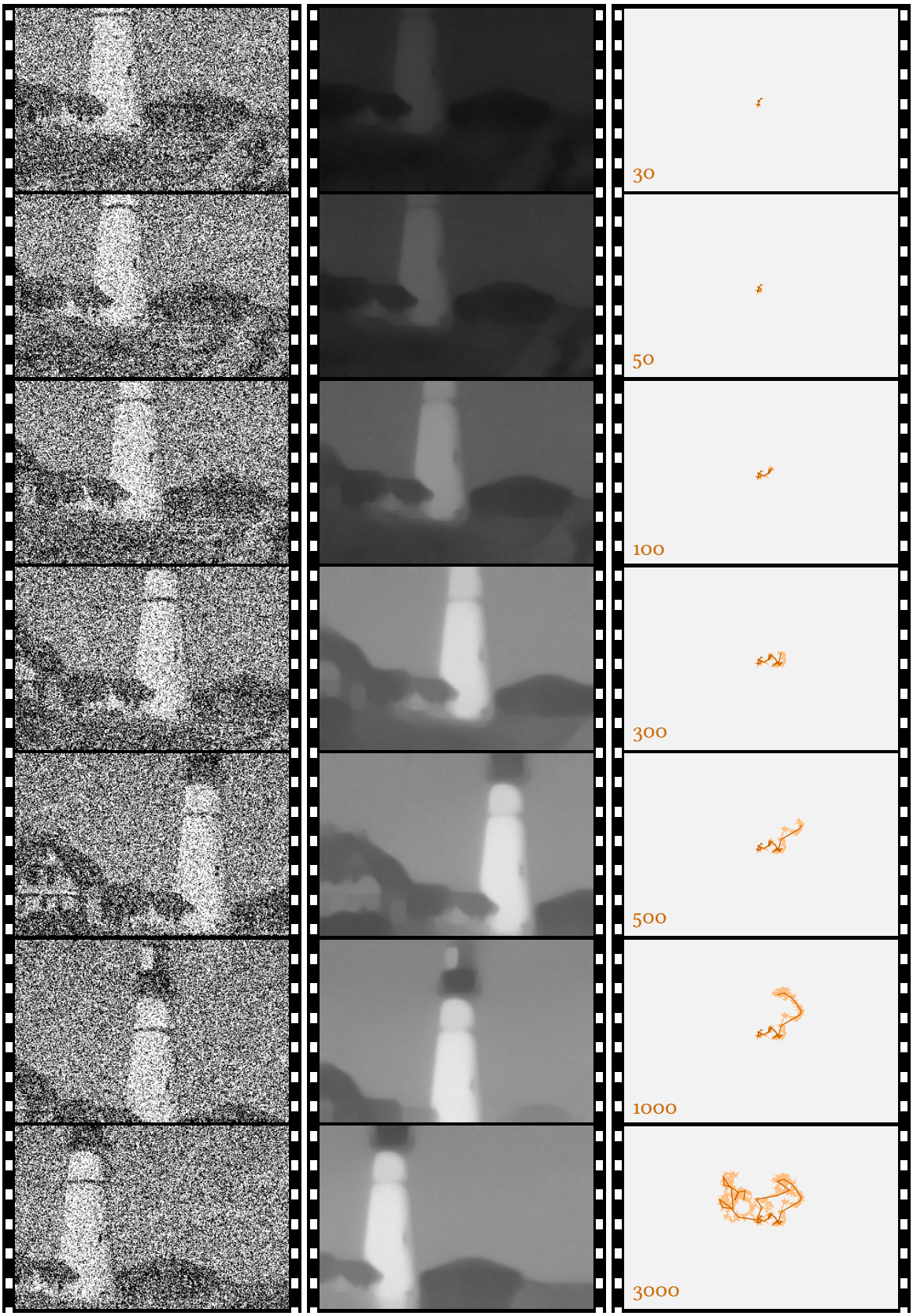
(a) Noisy (b) Reconstruction (c) Cumulative displacement

Figure 2: Square, POPD, approximately known displacement, $\rho = 0$.



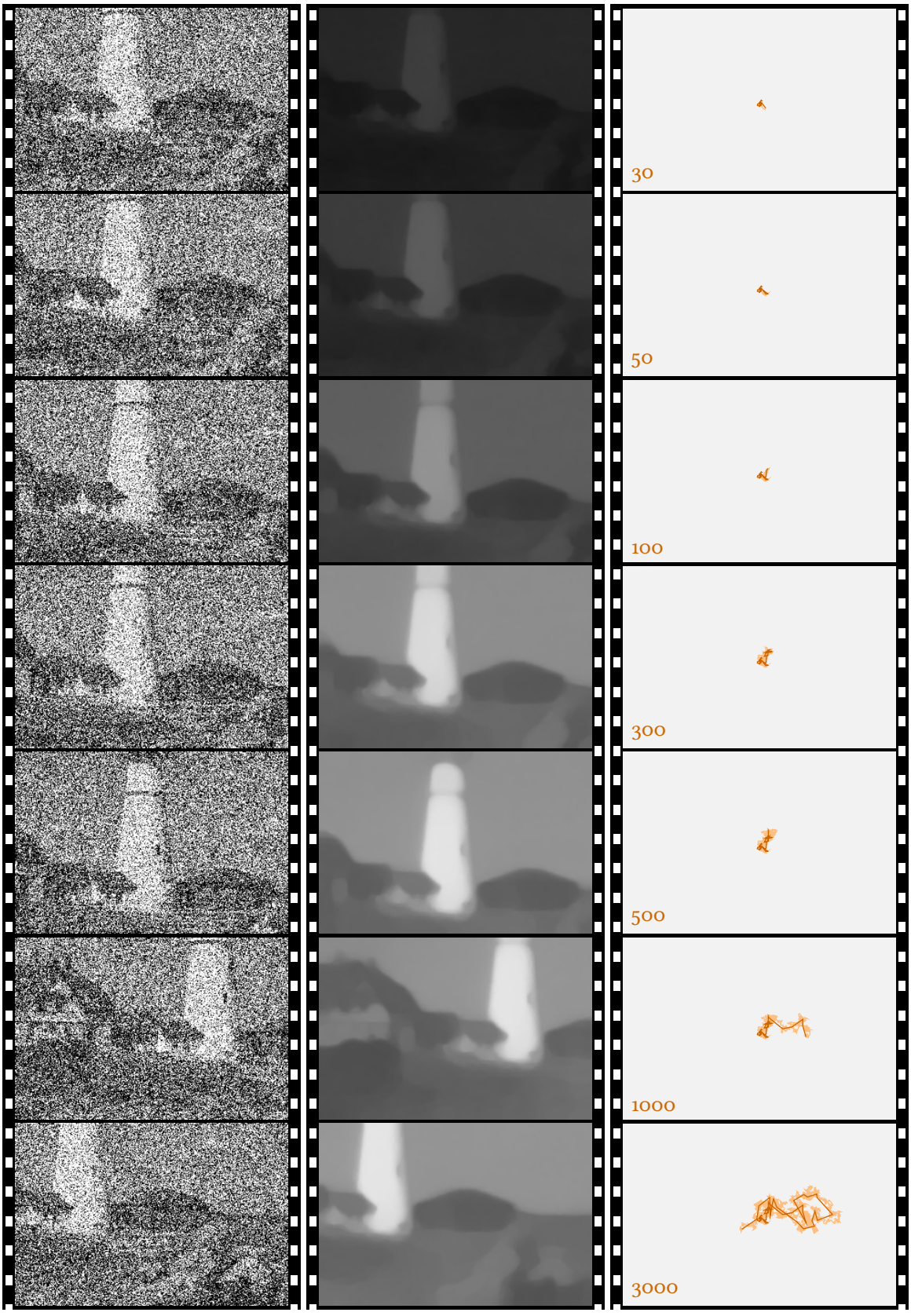
(a) Noisy (b) Reconstruction (c) Cumulative displacement

Figure 3: Lighthouse, POPD, approximately known displacement, $\rho = 0$.



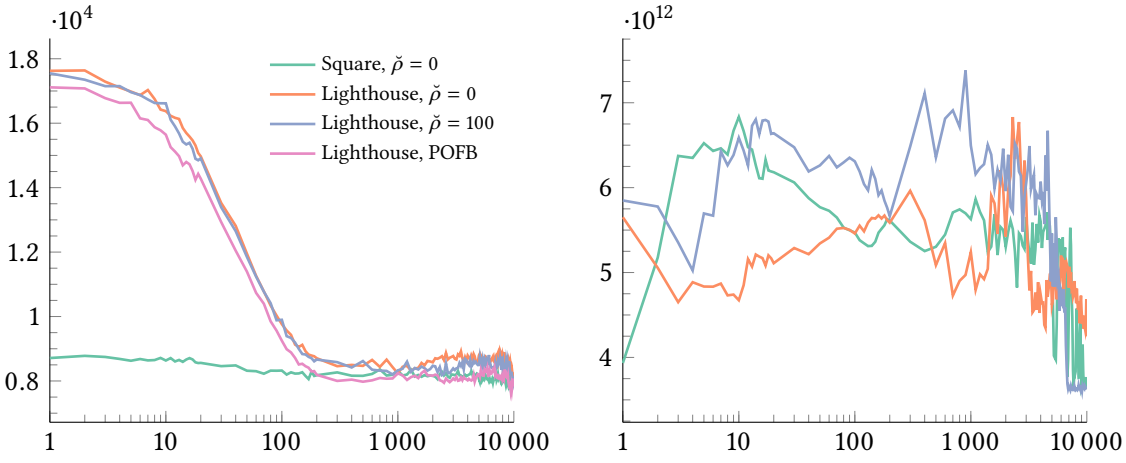
(a) Noisy (b) Reconstruction (c) Cumulative displacement

Figure 4: Lighthouse, POPD, approximately known displacement, $\rho = 100$.



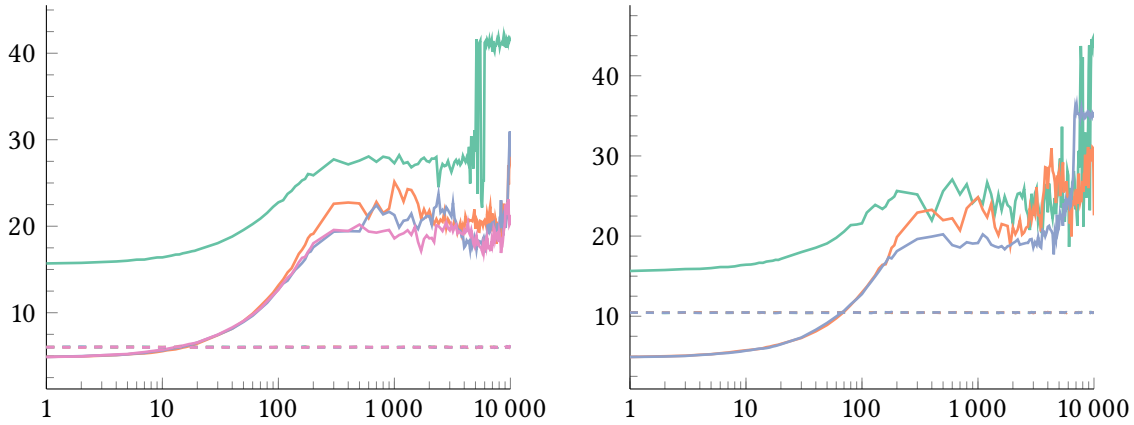
(a) Noisy (b) Reconstruction (c) Cumulative displacement

Figure 5: Lighthouse, POFB, approximately known displacement.



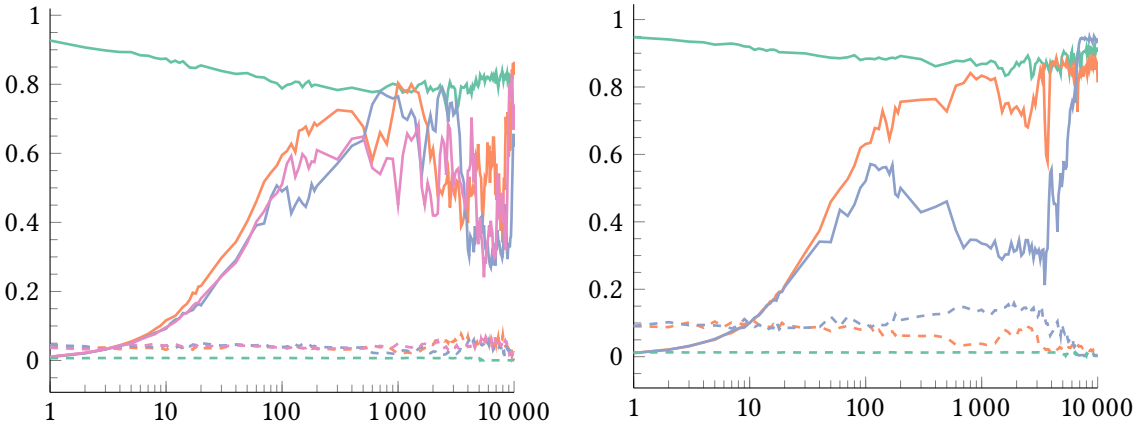
(a) Approximately known displacement (b) Unknown displacement

Figure 6: Iteration-wise objective values.



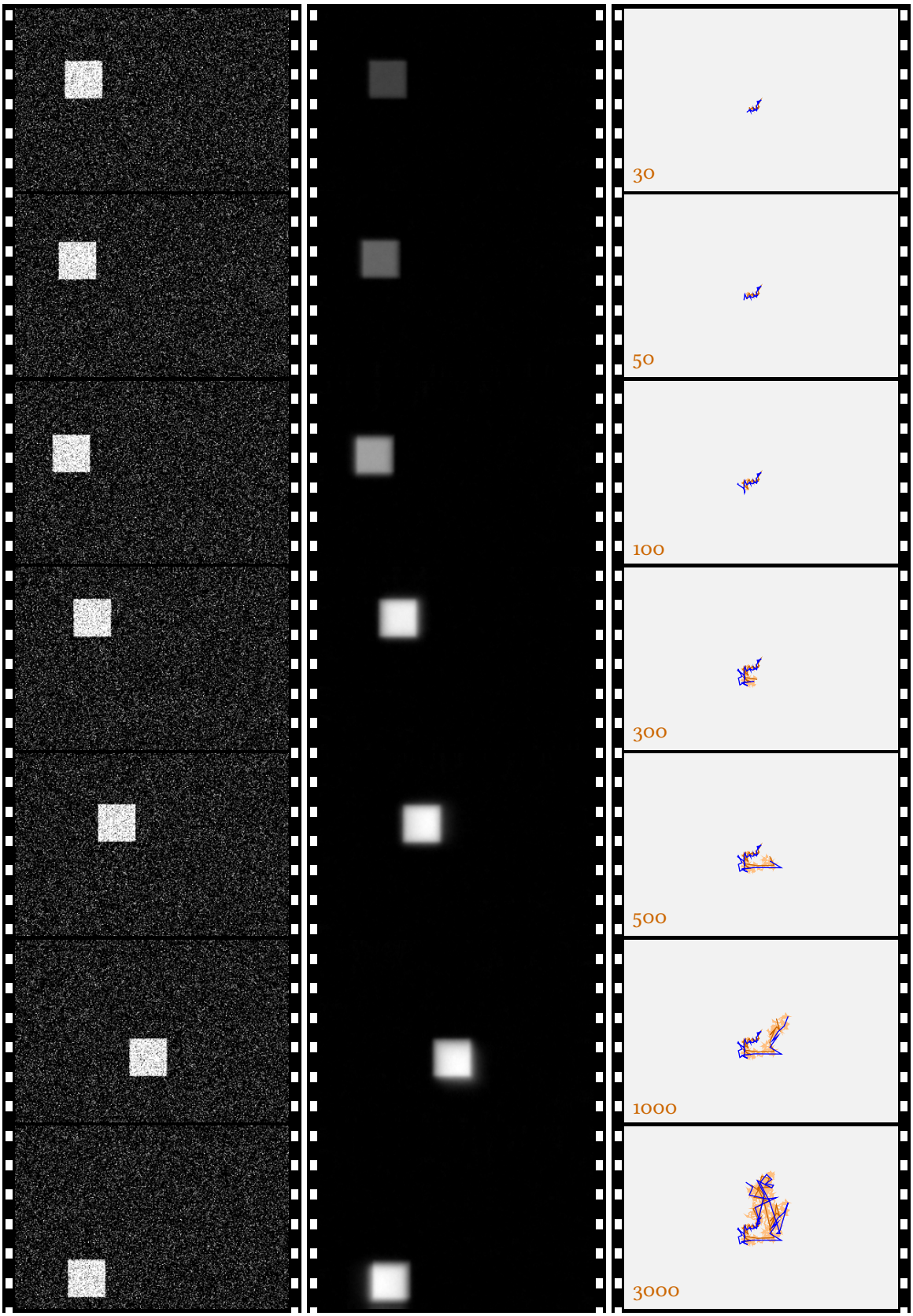
(a) Approximately known displacement (b) Unknown displacement

Figure 7: Iteration-wise PSNR. The dashed lines indicate the PSNR for the noisy data corresponding to the experiment of the solid line of the same colour. Legend in Figure 6a.



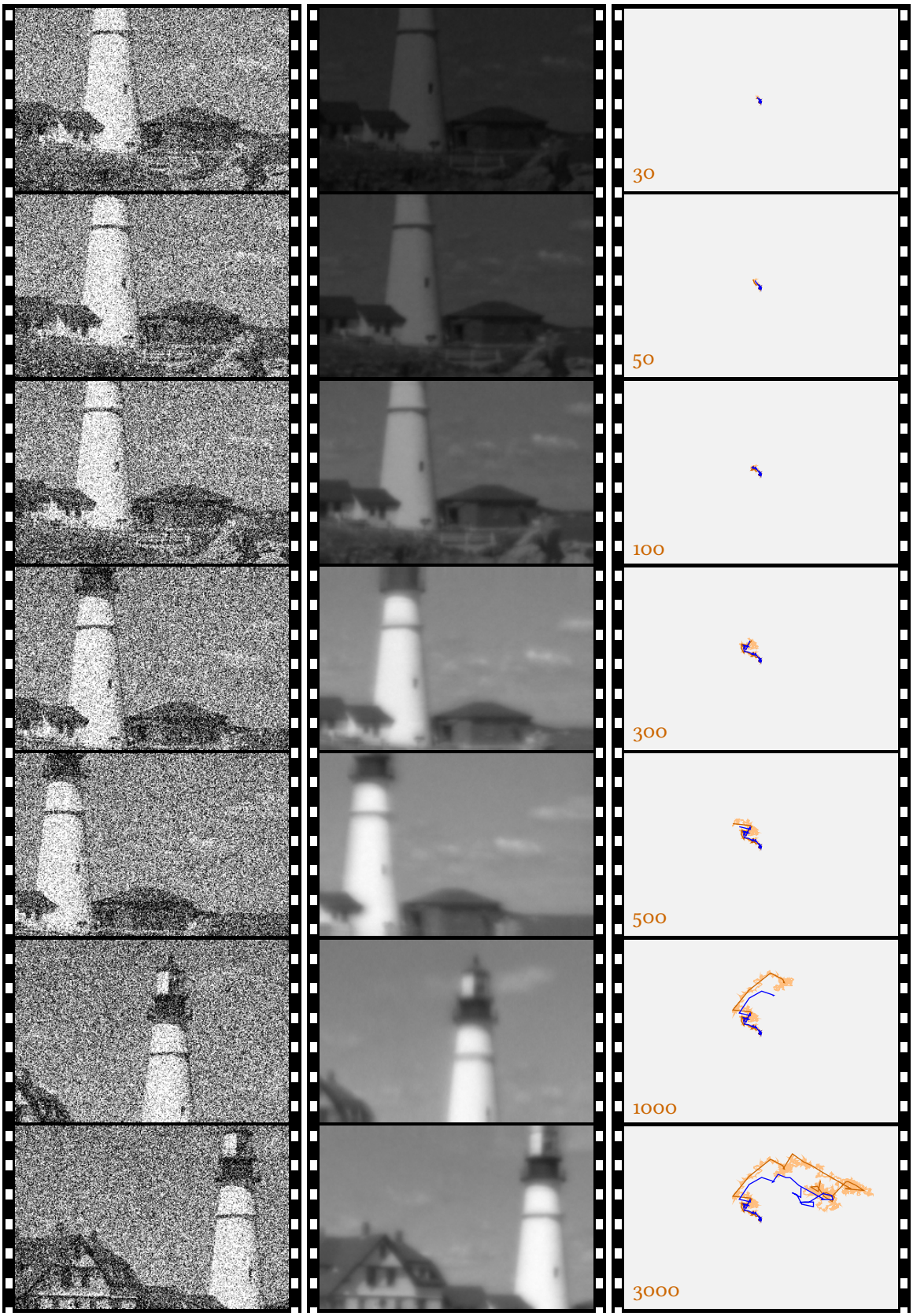
(a) Approximately known displacement (b) Unknown displacement

Figure 8: Iteration-wise SSIM. The dashed lines indicate the SSIM for the noisy data corresponding to the experiment of the solid line of the same colour. Legend in Figure 6a.



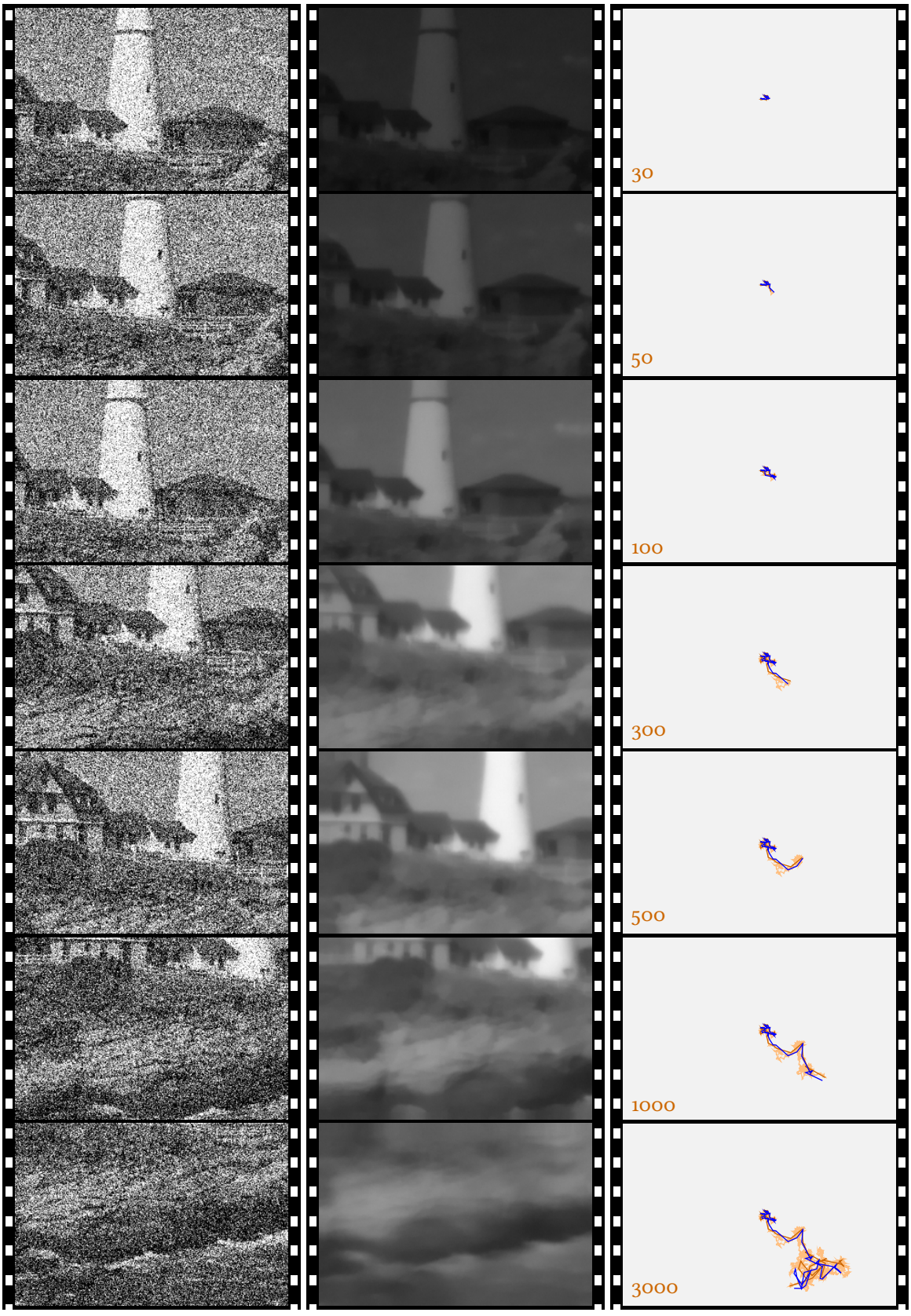
(a) Noisy (b) Reconstruction (c) Cumulative displacement

Figure 9: Square, POPD, unknown displacement, $\rho = 0$. The blue line in (c) indicates the estimated displacement field.



(a) Noisy (b) Reconstruction (c) Cumulative displacement

Figure 10: Lighthouse, POPD, unknown displacement, $\rho = 0$. The blue line in (c) indicates the estimated displacement field.



(a) Noisy (b) Reconstruction (c) Cumulative displacement

Figure 11: Lighthouse, POPD, unknown displacement, $\rho = 100$. The blue line in (c) indicates the estimated displacement field.