PRECONDITIONED PROXIMAL POINT METHODS IN HILBERT SPACES

Lectures for the 2018 Würzburg Winter School
"Modern methods in non-smooth optimisation"

Tuomo Valkonen

2018-04-09

CONTENTS

1 INTRODUCTION

These lectures are based on the articles [26, 30, 31], as well as the lecture notes [27, 29].

We assume that the reader is familiar with basic convex analysis, including in particular convex subdifferentials; an introduction may be found, for example, in [27, 29], and more in-depth details in [16, 24]. Before going forward, the reader may also wish to refer to Appendix A to refresh their mind on notation.

1.1 HOW TO DERIVE OPTIMISATION METHODS?

Let $f : X \to \mathbb{R}$ be convex, proper, and lower semicontinuous, on a Hilbert space X; we denote this $f \in \Gamma(X)$. We want to find a point $\hat{x}$ such that

$$f(\hat{x}) = \min_{x \in X} f(x). \tag{P}$$

This is of course characterised by

$$0 \in \partial f(\hat{x}).$$

This system is, however, in most interesting cases difficult to solve analytically. So let us try to derive numerical methods. One way of deriving numerical methods is to replace the original difficult objective with a simpler one whose minimisation provides improvement to the original objective.

Definition 1.1. A function $\widetilde{f}_{\bar{x}} : X \to \overline{\mathbb{R}}$ is a *surrogate objective* for $f : X \to \overline{\mathbb{R}}$ at $\bar{x}$ if $\widetilde{f}_{\bar{x}} \geq f$, and $\widetilde{f}_{\bar{x}}(\bar{x}) = f(\bar{x})$.

Starting with a point x^0, we would then minimise $\widetilde{f}_{x^0}$ to obtain a new point x^{i+1}. Through the properties of the surrogate objective, this will not increase the value of f. Hopefully it will provide significant improvement! Then we repeat the process, minimising $\widetilde{f}_{x^1}$, and so on.

1.2 GRADIENT DESCENT

What options are there for surrogate objectives, and which would be a good one? If f is (Fréchet) differentiable, one possibility is

$$\min_{x\in X} \widetilde{f}_{\bar{x}}(x) := f(\bar{x}) + \langle \nabla f(\bar{x}), x - \bar{x}\rangle + \frac{1}{2\tau}\|x - \bar{x}\|^2. \tag{1.1}$$

To show that $\widetilde{f}_{\bar{x}}$ is a surrogate function for suitable factors $\tau > 0$, we need the following definition:

Definition 1.2. Let $f : X \to \mathbb{R}$ be convex. We say that f is *L-smooth* if it is differentiable and

$$f(x') \le f(x) + \langle \nabla f(x), x' - x\rangle + \frac{L}{2}\|x' - x\|^2, \quad (x', x \in X). \tag{1.2}$$

In general $f(\bar{x}) = \widetilde{f}_{\bar{x}}(\bar{x})$. If f is L-smooth per Definition 1.2, and $L\tau \le 1$, then also $f \le \widetilde{f}_i$. Therefore, in this case, $\widetilde{f}_{\bar{x}}$ is a valid surrogate objective, and minimising $\widetilde{f}_{\bar{x}}$ will provide improvement to f as well.

The optimality condition $0 \in \partial \widetilde{f}_{x^i}(x)$ becomes

$$\nabla f(x^i) + \tau^{-1}(x - x^i) = 0. \tag{1.3}$$

This holds if $x^i = \hat{x}$ by taking also $x = \hat{x}$. Therefore, there is a direct correspondence between the solutions of the surrogate objective and the original. If $x^i \ne \hat{x}$, solving (1.3) for $x = x^{i+1}$, we get the rule

$$x^{i+1} = x^i - \tau \nabla f(x^i). \tag{GD}$$

This is known as the *gradient descent method*. In this context the quadratic term in (1.1) can be seen as a step length condition.

Will sequentially minimising $\widetilde{f}_{x^i}$ provide sufficient decrease in f such that we obtain convergence of $\{x^i\}$ to a minimiser $\hat{x}$ of f? A conventional way to do this is via Browder's fixed point theorem; see, e.g. [29]. In this course, we will introduce a different approach in Chapter 2.

1.3 THE PROXIMAL POINT METHOD

The gradient descent method is very basic, but often not very good. In particular, subgradient extensions of (GD) can have very slow convergence. Therefore we need alternative methods.

We now allow for general (possibly non-differentiable) convex functions $f : X \to \overline{\mathbb{R}}$, and replace the surrogate objective in (1.1) by another surrogate

$$\min_{x\in\mathbb{R}^n} \bar{f}_{\bar{x}}(x) := f(x) + \frac{1}{2\tau}\|x - \bar{x}\|^2. \tag{1.4}$$

In other words, we remove the linearisation, and try to minimise f directly with a step length condition. Again $\bar{f}_{\bar{x}}(\bar{x}) = f(\bar{x})$, and clearly $f_{\bar{x}} \geq f$. Therefore $\bar{f}_{\bar{x}}$ is a valid surrogate objective for f at $\bar{x}$. This time the optimality conditions for x minimising $\bar{f}_{x^i}$ are

$$0 \in \partial f(x) + \tau^{-1}(x - x^i). \tag{1.5}$$

If $x^i = \hat{x}$ for $\hat{x}$ a minimiser of the original objective f, then (1.5) is solved by $x = \hat{x}$, so again there is a direct correspondence between the solutions of the surrogate objective and the original.

The method based on solving (1.5) resp. (1.4) is known as the *proximal point method*. The step is the *backward step*, or the *implicit step*, since we cannot in general derive an explicit solution $x = x^{i+1}$, and try to go "back to x^i from x^{i+1}". However, *often, and especially in the context of splitting algorithms,* (1.5) *is easy to solve.* We will get back to this. By contrast, the gradient descent step (GD) is also known as the *forward step* or the *explicit step*, because we calculate $\nabla f(x^i)$ already at the current iterate, going "forward" from it.

Re-ordering as

$$x^i \in \tau\partial f(x^{i+1}) + x^{i+1},$$

the iteration resulting from the condition (1.5) may also be written as

$$x^{i+1} := (I + \tau\partial f)^{-1}(x^i), \tag{PP$_0$}$$

where the *proximal mapping*

$$\operatorname{prox}_{\tau\partial f} := (I + \tau\partial f)^{-1}$$

is the inverse of the set-valued map $A := I + \tau\partial f$, defined simply as

$$A^{-1}y := \{x \mid y \in Ax\}.$$

(Thus $y \in Ax \iff x \in A^{-1}y$.) As is evident from the expression

$$\operatorname{prox}_{\tau\partial f}(x) = \arg\min_{x'} f(x') + \frac{1}{2\tau}\|x' - x\|^2, \tag{1.6}$$

the proximal mapping is, in fact, single-valued.

Example 1.1. Let $f(x) = \|z-x\|_2^2/2$ for some $z \in \mathbb{R}^n$. By (1.6), we have $x' = \operatorname{prox}_{\tau\partial f}(x)$ if and only if $x \in \tau\partial f(x') + x'$. This gives the requirement $x = \tau(x'-z)+x'$. Consequently

$$\operatorname{prox}_{\tau\partial f}(x) = \frac{x+\tau z}{1+\tau}.$$

Example 1.2. Let $f(x) = \delta_{[-1,1]}$ on $\mathbb{R}$. Then by (1.6), $x' = \operatorname{prox}_{\tau\partial f}(x)$ if and only if $x \in \tau N_{[-1,1]}(x') + x'$. Since $z \in N_C(x')$ implies $\tau z \in N_C(x')$ for any convex set C and $\tau > 0$, this is to say $x \in x' + N_{[-1,1]}(x')$. Since

$$N_{[-1,1]}(x') = \begin{cases} [0,\infty), & x' = 1, \\ \{0\}, & x' \in (-1,1), \\ (-\infty, 0], & x' = 1, \\ \emptyset, & \text{otherwise}, \end{cases}$$

it is not difficult to verify that

$$x' = x \cdot \min\{1, 1/|x|\} = \begin{cases} 1, & x > 1, \\ x, & x \in [-1,1], \\ -1, & x < -1. \end{cases}$$

In other words, the proximal mapping is the (Euclidean) projection of x to $[-1,1]$. This is true in the general case $f(x) = \delta_C$, as is already evident from (1.6).

Exercise 1.1. *Calculate* $\operatorname{prox}_{\tau\partial f}$ *on* $\mathbb{R}^n$ *for*

(i) $f(x) = \delta_{\alpha B}(x)$, *where* B *is the unit ball and* $\alpha > 0$.

(ii) $f(x) = \alpha\|x\|_2$.

1.4 MONOTONE VARIATIONAL INCLUSIONS

The proximal point method (PP_0) readily generalises to solving for monotone $H : X \rightrightarrows X$ the *variational inclusion*

(MVI) $$0 \in H(x).$$

Definition 1.3. We recall that a set-valued map $H : X \rightrightarrows X$ is monotone, if

$$\langle q'-q, x'-x\rangle \geq 0 \quad ((q',y'),(x',x) \in \operatorname{Graph} H).$$

The method is simply

$$x^{i+1} := \operatorname{prox}_{\tau H}(x^i) = (I + \tau H)^{-1}(x^i). \tag{MPP}$$

The problem with (MPP) for any interesting H is that it will be just as difficult to solve as the original problem (MVI). There are however some ingenious ways to modify the step (MPP) to be cheap for specific problems with interesting H.

1.5 SADDLE POINT PROBLEMS

For some $g \in \Gamma(X)$, $f \in \Gamma(Y)$, and $K \in \mathcal{L}(X;Y)$, let us consider the problem

$$\min_x g(x) + f(Kx).$$

By writing f in terms of its convex conjugate f^*, we are led to

$$\min_x \max_y \; g(x) + \langle Kx, y\rangle - f^*(y), \tag{1.7}$$

Using the fact that $y \in \partial f(z)$ if and only if $z \in \partial f^*(y)$, which follows from f being convex, proper, and lower semicontinuous [see, e.g.. 16, 24, 29], the optimality conditions for this system can be seen to be

$$-K^*\hat{y} \in \partial g(\hat{x}), \quad \text{and} \quad K\hat{x} \in \partial f^*(\hat{y}).$$

These conditions may be encoded as $0 \in H(x,y)$ in terms of

$$H(x,y) := \begin{pmatrix} \partial g(x) + K^* y \\ \partial f^*(y) - Kx \end{pmatrix}. \tag{1.8}$$

In principle, we may therefore apply (MPP) to solve the saddle point problem (1.7). In practise we however need to work a little bit more, as the step (MPP) can rarely be given an explicit, easily solvable form.

Exercise 1.2. *With g and f convex, proper, and lower semicontinuous, verify that H given in (1.8) is a monotone operator as per Definition 1.3.*

1.6 DECOUPLING PRECONDITIONERS, THE PDHGM

However, there is a very effective primal–dual method for (1.7), that can be obtained from (MPP) with a small change. Let us first write out in explicit form the algorithm, known as the PDHGM (Primal–Dual Hybrid Gradient method, Modified) or the Chambolle–Pock method.

For parameters $\tau, \sigma > 0$, the primal variable x, and the dual variable y, we specifically iterate

$$x^{i+1} := (I + \tau \partial g)^{-1}(x^i - \tau K^* y^i), \tag{1.9a}$$

$$\bar{x}^{i+1} := 2x^{i+1} - x^i, \tag{1.9b}$$

$$y^{i+1} := (I + \sigma \partial f^*)^{-1}(y^i + \sigma K \bar{x}^{i+1}). \tag{1.9c}$$

The step (1.9a) is simply a proximal step for x in (1.7), keeping $y = y^i$ fixed. The step (1.9c) is likewise a proximal step for y in (1.7), keeping x fixed, not to x^i or x^{i+1} but to the *inertial variable* $\bar{x}^{i+1}$ defined in (1.9b). This may be visualised as a "heavy ball" version of x^{i+1} that has enough inertia to not get stuck in small bumps in the landscape.

With the general notation

$$u = (x, y),$$

the algorithm (1.9) may also be written in the *preconditioned* proximal point form

$$H(u^{i+1}) + M(u^{i+1} - u^i) \ni 0, \tag{1.10}$$

for the monotone operator H as in (1.8), and the preconditioning matrix

$$M := \begin{pmatrix} I/\tau & -K^* \\ -K & I/\sigma \end{pmatrix}.$$

Through the replacement of I by M in the basic proximal point iteration $u^{i+1} := (I+H)^{-1}(u^i)$, we thus have in (1.9a)–(1.9b) a proximal point method for which the steps can often be solved explicitly.

Theorem 1.1. *Let $f \in \Gamma(Y)$, $g \in \Gamma(X)$, and $K \in \mathcal{L}(X; Y)$. Choose $\tau, \sigma > 0$ such that $\tau\sigma\|K\|^2 < 1$. Let $u^* = (x^*, y^*)$ be a cluster point of the sequence of iterates $\{u^i\}$ generated by (1.9) for any starting point $u^0 = (x^0, y^0)$. Then u^* is a saddle point of (1.7).*

Proof. A saddle point $\widehat{u}$ satisfies $0 \in H(\widehat{u})$. Therefore by the monotonicity of H,

$$\langle H(u^{i+1}), u^{i+1} - \widehat{u} \rangle \geq 0.$$

Thus (1.10) gives

$$\langle M(u^{i+1} - u^i), u^{i+1} - \widehat{u} \rangle \leq 0. \tag{1.11}$$

With the notation $\|x\|_M := \sqrt{\langle Mx, x \rangle}$, we have

$$\langle M(u^{i+1} - u^i), u^{i+1} - \widehat{u} \rangle = \frac{1}{2}\|u^{i+1} - u^i\|_M^2 - \frac{1}{2}\|u^i - \widehat{u}\|_M^2 + \frac{1}{2}\|u^{i+1} - \widehat{u}\|_M^2.$$

Now (1.11) shows that

$$\frac{1}{2}\|u^{i+1}-\widehat{u}\|_M^2+\frac{1}{2}\|u^{i+1}-u^i\|_M^2\leq\frac{1}{2}\|u^i-\widehat{u}\|_M^2. \tag{1.12}$$

Summing (1.12) over $i=0,\dots,N-1$ shows that

$$\frac{1}{2}\|u^{N}-\widehat{u}\|_M^2+\sum_{i=0}^{N-1}\frac{1}{2}\|u^{i+1}-u^i\|_M^2\leq\frac{1}{2}\|u^0-\widehat{u}\|_M^2. \tag{1.13}$$

Now, the condition $\tau\sigma\|K\|^2<1$ ensures that $\|u\|_M^2\geq\theta\|u\|^2$ for some $\theta>0$. Therefore (1.13) shows that $\|u^{i+1}-u^i\|\to 0$, and that $\{u^i\}_{i\in\mathbb{N}}$ is bounded. Therefore, every subsequence $\{u^{i_j}\}_{j\in\mathbb{N}}$ has a further subsequence that converges to some point u^* satisfying $0\in H(u^*)$. In particular, every cluster point is a saddle point. □

Exercise 1.3. *Using Opial's lemma below, show that there is, in fact, only one cluster point. Show, therefore, that the whole sequence of iterates converges to a saddle point.*

The earliest version of the next lemma, required for Exercise 1.3, is contained in the proof of [22, Theorem 1]. A more complete statement can be found as [4, Lemma 6].

Lemma 1.2 (Opial's lemma). *On a Hilbert space X, let $\hat{X}\subset X$ be closed and convex, and $\{x^i\}_{i\in\mathbb{N}}\subset X$. If the following conditions hold, then $x^i\rightharpoonup x^*$ weakly in X for some $x^*\in\hat{X}$:*

(i) $i\mapsto\|x^i-x^*\|$ is non-increasing for all $x^*\in\hat{X}$.

(ii) All weak limit points of $\{x^i\}_{i\in\mathbb{N}}$ belong to $\hat{X}$.

The property (i) of Opial's lemma has a name worth remembering:

Definition 1.4 (Féjer monotonicity). Given a non-empty subset $\hat{X}\subset X$, a sequence $\{u^i\}_{i\in\mathbb{N}}$ is *Féjer monotone* with respect to $\hat{X}$ if

$$\|u^{i+1}-u\|\leq\|u^i-u\|\quad(i\in\mathbb{N};u\in\hat{X}).$$

We refer to [2] for more information on this property.

1.7 TRICKS OF THE TRADE

Example 1.3 (Dualisation trick for hard-to-invert forward operators). As we have seen in Example 1.1, the proximal mapping of $g(x)=\|z-x\|_2^2/2$ is easy to calculate. But what about $g(x)=\|z-Ax\|_2^2/2$ for some $A\in\mathbb{R}^{k\times n}$ and $z\in\mathbb{R}^k$? Unless A is unitary (i.e., $A^*A=I$, such as a Fourier transform), the computation of $\operatorname{prox}_{\tau\partial f}$ will generally

require a costly matrix inversion. However, we can also use the *dualisation trick*

$$g(x) = \sup_{\lambda \in \mathbb{R}^k} \langle Ax - z, \lambda \rangle - \frac{1}{2}\|\lambda\|^2,$$

and replace the saddle point problem

$$\min_x \max_y \; g(x) + \langle Kx, y \rangle - f^*(y)$$

by

$$\min_x \max_{\widetilde{y}} \; \widetilde{g}(x) + \langle \widetilde{K}x, \widetilde{y} \rangle - \widetilde{f}^*(\widetilde{y}),$$

where $\widetilde{y} = (y, \lambda)$ and the mappings

$$\widetilde{g}(x) = 0, \quad \widetilde{f}^*(\widetilde{y}) = f^*(y) + \frac{1}{2}\|\lambda\|^2 + \langle z, \lambda \rangle, \quad \text{and} \quad \widetilde{K}x = (Kx, Ax).$$

1.8 FORWARD–BACKWARD SPLITTING

Let us consider the minimisation of the composite objective

$$\min_{x \in \mathbb{R}^n} h(x) := g(x) + f(x), \tag{1.14}$$

where g is smooth, but f possibly non-smooth. We may write the optimality conditions as

$$0 \in \nabla g(x) + \partial f(x).$$

We can rewrite this as

$$\tau^{-1}x - \nabla g(x) \in \tau^{-1}x + \partial f(x),$$

or

$$x = (I + \tau \partial f)^{-1}(x - \tau \nabla g(x)).$$

This gives the iteration

$$x^{i+1} = \operatorname{prox}_{\tau \partial f}(x^i - \tau \nabla g(x^i)). \tag{FB}$$

In other words, we do a gradient/forward step with respect to g, and a proximal/backward step with respect to f. The resulting method is known as *forward–backward splitting*. Particular instances include the so-called *iterative soft-thresholding (IST)* algorithm for Lasso, with $\operatorname{prox}_{\tau \partial |\cdot|}$ known as the iterative soft-thresholding operator.

Exercise 1.4. *Express forward–backward splitting in terms of a surrogate objective.*

2 PRECONDITIONED PROXIMAL POINT METHODS

Our overall wish is to find some $\widehat{u} \in U$, on a Hilbert space U, solving for a given set-valued map $H : U \rightrightarrows U$ the variational inclusion

$$0 \in H(\widehat{u}). \tag{2.1}$$

Our strategy towards finding a solution $\widehat{u}$ is to introduce an arbitrary non-linear iteration-dependent *preconditioner* $V_{i+1} : U \to U$ and a *step length operator* $W_{i+1} \in \mathcal{L}(U;U)$. With these, we define the generalised proximal point method, which on each iteration $i \in \mathbb{N}$ solves u^{i+1} from

$$0 \in W_{i+1}H(u^{i+1}) + V_{i+1}(u^{i+1}). \tag{PP}$$

We assume that V_{i+1} splits into $M_{i+1} \in \mathcal{L}(U;U)$, and $V'_{i+1} : U \to U$ as

$$V_{i+1}(u) = V'_{i+1}(u) + M_{i+1}(u - u^i). \tag{2.2}$$

In contrast to (1.5) or (1.10), we place the step lengths at the front of H instead of inverted in M_{i+1}, in order to allow zero step lengths, as we will later discuss in Section 3.4.

Example 2.1 (The basic proximal point method). To obtain the basic proximal point method (PP$_0$, p.5), we set $W_{i+1} := \tau I$, $M_{i+1} := I$, $V'_{i+1} \equiv 0$, and $H := \partial f$.

Example 2.2 (The PDHGM or Chambolle–Pock method). To obtain the PDHGM (1.9) for saddle point problems, we take H as in (1.8), and set

$$W_{i+1} := \begin{pmatrix} \tau I & 0 \\ 0 & \sigma I \end{pmatrix}, \quad M_{i+1} := \begin{pmatrix} I & -\tau K^* \\ -\sigma K & I \end{pmatrix}, \quad \text{and} \quad V'_{i+1} \equiv 0.$$

Indeed, (PP) is (1.10) multiplied by W_{i+1}.

We analyse (PP) by applying a *testing operator* $Z_{i+1} \in \mathcal{L}(U;U)$, following the ideas introduced in [32]. The product $Z_{i+1}M_{i+1}$ with the linear part of the preconditioner, forms a "metric" $\|\cdot\|^2_{Z_{i+1}M_{i+1}}$, which will, as we soon demonstrate, be an indicator of convergence rates.

2.1 A GENERAL CONVERGENCE RESULT

Adelante! We go straight ahead with our main convergence result, an almost trivial little theorem, on which everything that follows is based on.

Theorem 2.1. *On a Hilbert space U, let $H : U \rightrightarrows U$, and $W_{i+1}, M_{i+1}, Z_{i+1} \in \mathcal{L}(U;U)$, as well as $V'_{i+1} : U \to U$ for $i \in \mathbb{N}$. Suppose* (PP) *is solvable for V_{i+1} as in* (2.2)*, and denote the iterates by $\{u^i\}_{i\in\mathbb{N}}$. If $Z_{i+1}M_{i+1}$ is self-adjoint, and*

$$\text{(CI)}\quad \frac{1}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1}-\widehat{u}\|^2_{Z_{i+1}M_{i+1}-Z_{i+2}M_{i+2}} + \langle W_{i+1}H(u^{i+1}) + V'_{i+1}(u^{i+1}), u^{i+1}-\widehat{u}\rangle_{Z_{i+1}} \geq -\Delta_{i+1}(\widehat{u})$$

for all $i \in \mathbb{N}$ and some $\widehat{u} \in U$, then

$$\text{(DI)}\qquad \frac{1}{2}\|u^N-\widehat{u}\|^2_{Z_{N+1}M_{N+1}} \leq \frac{1}{2}\|u^0-\widehat{u}\|^2_{Z_1M_1} + \sum_{i=0}^{N-1}\Delta_{i+1}(\widehat{u}) \quad (N \geq 1).$$

Proof. Inserting (PP) into (CI), we obtain

$$\text{(2.3)}\quad \frac{1}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1}-\widehat{u}\|^2_{Z_{i+1}M_{i+1}-Z_{i+2}M_{i+2}} - \langle u^{i+1}-u^i, u^{i+1}-\widehat{u}\rangle_{Z_{i+1}M_{i+1}} \geq -\Delta_{i+1}(\widehat{u}).$$

We recall for general self-adjoint M the three-point formula

$$\text{(2.4)}\qquad \langle u^{i+1}-u^i, u^{i+1}-\widehat{u}\rangle_M = \frac{1}{2}\|u^{i+1}-u^i\|^2_M - \frac{1}{2}\|u^i-\widehat{u}\|^2_M + \frac{1}{2}\|u^{i+1}-\widehat{u}\|^2_M.$$

Using this with $M = Z_{i+1}M_{i+1}$, we rewrite (2.3) as

$$\text{(2.5)}\qquad \frac{1}{2}\|u^i-\widehat{u}\|^2_{Z_{i+1}M_{i+1}} - \frac{1}{2}\|u^{i+1}-\widehat{u}\|^2_{Z_{i+2}M_{i+2}} \geq -\Delta_{i+1}(\widehat{u}).$$

Summing this over $i = 0, \dots, N-1$, we obtain (DI). □

Remark 2.2 (Quantitative Féjer monotonicity). *If $\Delta_{i+1}(\widehat{u}) \equiv 0$, the inequality* (2.5) *is a quantitative or* variable-metric *version of Féjer monotonicity of Definition 1.4 with respect to $C = \{\widehat{u}\}$.*

Immediately we obtain the following:

Corollary 2.3 (Convergence with a rate). *Suppose* (DI) *holds with $\Delta_{i+1}(\widehat{u}) \leq 0$, and that $Z_{N+1}M_{N+1} \geq \mu(N)I$. Then $\|u^N-\widehat{u}\|^2 \to 0$ at the rate $O(1/\mu(N))$.*

For weak convergence, when we cannot make $Z_{N+1}M_{N+1}$ grow fast, we need to do a little bit additional technical work. For this we need a few concepts form the convergence of sets, and the continuity of set-valued maps. More details can be found in, e.g., [25].

2.2 INTERLUDE: SET CONVERGENCE AND MAXIMAL MONOTONE OPERATORS

Definition 2.1. Let $\{A^i\}_{i=1}^\infty$ be a sequence of subsets of X. The *(strong, resp. weak) outer limit* of the sequence is the set $\limsup_{i\to\infty} A^i \subset X$ of all $x^* \in X$ such that there exist a subsequence $\{i_k\}_{k=1}^\infty$ and $x^k \in A^{i_k}$ with $x^k \to x^*$ (strongly, resp. weakly).

Definition 2.2. A set-valued map $H : U \rightrightarrows U$ is weak-to-strong (resp. strong-to-strong) outer semicontinuous if $u^i \rightharpoonup u$ (resp. $u^i \to u$) implies $\limsup_{i\to\infty} H(u^i) \subset H(u)$.

It is well-known that convex subdifferentials $H = \partial f$ are maximal monotone by the following definition [see 2, 24]. They are also weak-to-strong outer semicontinous, as well as the opposite [see 2, Proposition 16.26 & Proposition 20.33].

Definition 2.3. A monotone operator $H : X \rightrightarrows X$ is *maximal* if there does not exist a monotone operator $T : X \rightrightarrows X$ with $\operatorname{Graph} H \subsetneq \operatorname{Graph} T$.

Lemma 2.4. *Let $H : U \rightrightarrows U$ be maximal monotone on a Hilbert space U. Then H is is weak-to-strong outer semicontinuous: for any sequence $\{u^i\}_{i\in\mathbb{N}}$, and any $z^i \in H(u^i)$ such that $u^i \rightharpoonup u$ weakly, and $z^i \to z$ strongly, we have $z \in H(u)$.*

Proof. By monotonicity, for any $u' \in U$ and $z' \in U$ holds $D_i := \langle u' - u^i, z' - z^i\rangle \geq 0$. Since a weakly convergent sequence is bounded, we have $D_i \geq \langle u' - u^i, z' - z\rangle - C\|z - z^i\|$ for some $C > 0$ independent of i. Taking the limit, we therefore have $\langle u' - u, z' - z\rangle \geq 0$. If we had $z \notin H(u)$, this would contradict that H is maximal. □

2.3 ADDITIONAL CONDITIONS FOR WEAK CONVERGENCE

Corollary 2.5 (Weak convergence). *Suppose $Z_iM_i = Z_0M_0 \geq 0$ is self-adjoint, and that the iterates of* (PP, p.11) *satisfy* (CI) *with $\Delta_{i+1}(\widehat{u}) \leq -\frac{\delta}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}$ for all $\widehat{u} \in H^{-1}(0)$ and some $\delta > 0$. If*

(CL) $Z_{i+1}M_{i+1}(u^{i+1}-u^i) \to 0$ *and* $u^{i_k} \rightharpoonup u \implies \limsup_{k\to\infty} W_{i_k+1}H(u^{i_k})+V'_{i_k+1}(u^{i_k}) \subset W_*H(u)$

for some non-singular $W_* \in \mathcal{L}(U;U)$, then $Z_0M_0(u^i - u^*) \rightharpoonup 0$ weakly in U for some $u^* \in H^{-1}(0)$.

The condition (CL) is clarified by the following corollary that fixes W_{i+1}.

Corollary 2.6 (Weak convergence, fixed steps). *Suppose $Z_iM_i = Z_0M_0 \geq 0$ is self-adjoint, $W_{i+1} \equiv W$, and that the iterates of* (PP, p.11) *satisfy* (CI) *with $\Delta_{i+1}(\widehat{u}) \leq -\frac{\delta}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}$ for all $\widehat{u} \in H^{-1}(0)$ and some $\delta > 0$. If H is weak-to-strong outer semicontinuous, and*

$$Z_{i+1}M_{i+1}(u^{i+1} - u^i) \to 0 \implies \limsup_{k\to\infty} V'_{i+1}(u^{i+1}) \to 0,$$

then $Z_0M_0(u^i - u^*) \rightharpoonup 0$ weakly in U for some $u^* \in H^{-1}(0)$.

Proof of Corollary 2.6. We apply Theorem 2.1 on any $\widehat{u} \in U'$. Since $Z_{i+1}M_{i+1} = Z_{i+2}M_{i+2}$, it is easy to see that (CI) and consequently by the theorem, (DI) holds for all $\widehat{u} \in U' := \operatorname{cl}\operatorname{conv}\hat{U}$.

For the rest of the proof, we use Opial's lemma (Lemma 1.2). Using $\Delta_{i+1}(\widehat{u}) \le -\frac{\delta}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}$, we deduce from (DI, p.12) that $A^{1/2}(u^{i+1} - u^i) \to 0$ for $A := Z_0M_0 = Z_{i+1}M_{i+1}$. By (PP, p.11) and (CL), any weak limit point u^* of a subsequence of the sequence $\{u^i\}_{i\in\mathbb{N}}$ then satisfies $u^* \in \hat{U} \subset U'$. This verifies condition (ii) of the lemma for $x^i := A^{1/2}u^i$ and $X' := A^{1/2}\hat{U}$ on $X := A^{1/2}U \subset U$. Applied with $N = 1$ and u^i in place of u^0, (DI) shows condition (i) of the lemma. Thus $x^i \rightharpoonup x^* \in \hat{X}$. But $x^* = A^{1/2}u^*$ for some $u^* \in \hat{U}$. Thus $A(u^i - u^*) \rightharpoonup 0$. This implies $Z_0M_0(u^i - u^*) \rightharpoonup 0$ weakly. □

2.4 EXAMPLES

We now look at several concrete examples.

Example 2.3 (The proximal point method). Take $M_i = I$, $V_i' = 0$, and $W_{i+1} = \tau_i I$ for some $\tau_i > 0$. (Recall Example 2.1.) Then (PP, p.11) is the standard proximal point method $x^{i+1} := (I + \tau_i H)^{-1}(x^i)$. If H is maximal monotone, $\{u^i\}_{i\in\mathbb{N}}$ converges weakly to some $u^* \in H^{-1}(0)$.

Verification. We take $Z_{i+1} = \phi_i I$ for some $\phi_i > 0$. As long as $\phi_i \ge \phi_{i+1}$, the monotonicity of H clearly shows (CI, p.12) with $\Delta_{i+1}(\widehat{u}) = -\frac{\phi_i}{2}\|u^{i+1} - u^i\|^2$. Using the maximal monotonicity, Minty's theorem [e.g., 2, Theorem 21.1] guarantees the solvability of (PP, p.11). Thus the conditions of Theorem 2.1 are satisfied. Maximal monotonicity also guarantees that H is weak-to-strong outer semicontinuous; see Lemma 2.4. This establishes (CL). Taking $\phi_i \equiv \phi_0$ for constant $\phi_0 > 0$, so that $Z_{i+1}M_{i+1} = Z_0M_0 = \phi_0 I$, it remains to refer to Corollary 2.6. □

Example 2.4 (Acceleration and linear convergence of the proximal point method). Continuing from Example 2.3, suppose H is strongly monotone. Then $\langle H(u^{i+1}) - H(\widehat{u}), u^{i+1} - \widehat{u}\rangle \ge \gamma\|u^{i+1} - \widehat{u}\|^2$ for some $\gamma > 0$, so (CI, p.12) continues to hold with $\Delta_{i+1}(\widehat{u}) = -\frac{\phi_i}{2}\|u^{i+1} - u^i\|^2$ if $\phi_i(1 + 2\gamma\tau_i) \ge \phi_{i+1}$. This is the case for $\tau_{i+1} := \tau_i/\sqrt{1 + 2\gamma\tau_i}$, and $\phi_{i+1} := 1/\tau_{i+1}^2$. The testing variable ϕ_N is of the order $\Theta(N^2)$ [5, 32], so we get convergence of $\|u^N - \widehat{u}\|^2$ to zero at the rate $O(1/N^2)$ from Theorem 2.1 and Corollary 2.3. Alternatively, if we keep $\tau_i = \tau_0$ constant, ϕ_N becomes exponential, so we get linear convergence.

The next lemma starts our analysis of gradient descent:

Lemma 2.7. *Let $H = \nabla G$ for $G \in \Gamma(X)$ such that ∇G is L-Lipschitz. Take $M_{i+1} \equiv I$ and $V_{i+1}'(u) := \tau_i(\nabla G(u^i) - \nabla G(u))$ with $W_{i+1} = \tau_i I$ as well as $Z_{i+1} \equiv \phi_i I$ for some $\tau_i, \phi_i > 0$. Then* (CI, p.12) *holds if*

(i) $\phi_i = \phi$ *is constant,* $\tau_i L < 2$*, and* $\Delta_{i+1}(\widehat{u}) := -\phi_i(1 - \tau_i L/2)\|u^{i+1} - u^i\|^2/2$.

If G *is strongly convex with factor* $\gamma > 0$*, alternatively:*

(ii) $\tau_0 L^2 < \gamma$, $\phi_{i+1} := \phi_i + \phi_i\tau_i(\gamma - \tau_i L^2)$, $\tau_i := \phi_i^{-1/2}$ *or* $\tau_i := \tau_0$*, and* $\Delta_{i+1}(\widehat{u}) = 0$.

Proof. We will instead of (CI, p.12) prove the following more general condition: for some $\Delta_{i+1}(u^*; u)$ at all $u, u^* \in U$ holds

$$\frac{1}{2}\|u - u^i\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u - u^*\|^2_{Z_{i+1}M_{i+1} - Z_{i+2}M_{i+2}} + \langle W_{i+1}(H(u) - H(u^*)) + V'_{i+1}(u), u - u^*\rangle_{Z_{i+1}} \geq -\Delta_{i+1}(u^*; u). \tag{2.6}$$

We start by expanding (2.6) as

$$\frac{\phi_i}{2}\|u - u^i\|^2 + \frac{\phi_i - \phi_{i+1}}{2}\|u - u^*\|^2 + \phi_i\tau_i\langle \nabla G(u^i) - \nabla G(u^*), u - u^*\rangle \geq -\Delta_{i+1}(u^*; u). \tag{2.7}$$

(i) Lipschitz gradient implies L^{-1}-co-coercivity (see Exercise 2.1 below)

$$\langle \nabla G(u') - \nabla G(u), u' - u\rangle \geq L^{-1}\|\nabla G(u') - \nabla G(u)\|^2 \quad \text{for all} \quad u, u'. \tag{2.8}$$

Now (2.7) follows after we use (2.8) and Cauchy's inequality to estimate

$$\langle \nabla G(u^i) - \nabla G(u^*), u - u^*\rangle = \langle \nabla G(u^i) - \nabla G(u^*), u^i - u^*\rangle + \langle \nabla G(u^i) - \nabla G(u^*), u - u^i\rangle \geq -\frac{L}{4}\|u - u^i\|^2. \tag{2.9}$$

(ii) We estimate

$$\langle \nabla G(u^i) - \nabla G(u^*), u - u^*\rangle = \langle \nabla G(u) - \nabla G(u^*), u - u^*\rangle + \langle \nabla G(u^i) - \nabla G(u), u - u^*\rangle \geq \frac{\gamma}{2}\|u - u^*\|^2 - \frac{1}{2\tau_i}\|u - u^i\|^2 - \frac{\tau_i L^2}{2}\|u - u^*\|^2.$$

Inserting this into (2.7), we see that (2.6) holds with $\Delta_{i+1}(u^*; u) = 0$ if

$$\phi_i + \phi_i\tau_i(\gamma - \tau_i L^2) \geq \phi_{i+1}. \tag{2.10}$$

Clearly our two alternative choices of $\{\tau_i\}_{i\in\mathbb{N}}$ are non-increasing. Therefore, (2.10) follows from the initialisation condition $\tau_0 L^2 < \gamma$ and the update rule $\phi_{i+1} := \phi_i + \phi_i\tau_i(\gamma - \tau_i L^2)$. □

Exercise 2.1. *Prove that* ∇G *being* L*-Lipschitz for a convex function* G *implies the* L^{-1}*-co-coercivity* (2.8).

Example 2.5 (Gradient descent). Taking $\tau_i = \tau$ constant in Lemma 2.7, (PP, p.11) reads

$$0 = \tau \nabla G(u^i) + u^{i+1} - u^i.$$

This is the gradient descent method. Direct application of Lemma 2.7(i) with $u = u^{i+1}$ and $u^* = \widehat{u}$ together with Theorem 2.1 and Corollary 2.6 now verifies the well-known weak convergence of the method when $\tau L < 2$.

Observe that $V_{i+1} = \nabla Q_{i+1}$ for

$$Q_{i+1}(u) := \frac{1}{2}\|u - u^i\|^2 + \tau \left[G(u^i) + \langle \nabla G(u^i), u - u^i\rangle - G(u)\right].$$

As we have already seen, each step of (PP) therefore minimises the *surrogate objective*

$$u \mapsto G(u) + \tau^{-1} Q_{i+1}(u). \tag{2.11}$$

The function Q_{i+1} on one hand penalises long steps, and on the other hand allows longer steps when the local linearisation error is large.

Example 2.6 (Forward–backward splitting). Let $H = \partial G$, where $G = G_0 + J$ for $G, F \in \Gamma(X)$ with ∇J Lipschitz. Taking M_{i+1}, W_{i+1}, and V'_{i+1} as in Example 2.5, (PP, p.11) becomes

$$0 \in \tau_i \partial G_0(u^{i+1}) + \tau_i \nabla J(u^i) + u^{i+1} - u^i.$$

This is the forward–backward splitting method

$$u^{i+1} := (I + \tau_i \partial G_0)^{-1}(u^i - \tau_i \nabla J(u^i)).$$

The method converges when the gradient descent of Example 2.5 applied to J converges.

Exercise 2.2. *Prove the convergence claims of Example 2.6 for forward–backward splitting.*

Hint: *Use (2.6) for a suitable choice of u^*.*

3 SADDLE POINT PROBLEMS AND PRIMAL–DUAL METHODS

3.1 THE THEORY, SPECIALISED

With $K \in \mathcal{L}(X;Y)$, $G = G_0 + J \in \Gamma(X)$, and $F^* \in \Gamma(Y)$ on Hilbert spaces X and Y, we now wish to solve the saddle point or min–max problem (1.7). We recall that the first-order necessary optimality conditions can be written

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

Setting $U := X \times Y$ and introducing the variable splitting notation $u = (x, y)$, $\widehat{u} = (\widehat{x}, \widehat{y})$, etc., this can succinctly be written as $0 \in H(\widehat{u})$ in terms of the operator H from (1.8). We recall this to be

$$H(u) := \begin{pmatrix} \partial G(x) + K^* y \\ \partial F^*(y) - Kx \end{pmatrix}. \tag{3.1}$$

In this section, concentrating on this specific H, we specialise the theory of Section 2.1 to saddle point problems. Throughout, for some primal and dual step length and testing parameters $\tau_i, \phi_i > 0, \psi_{i+1} > 0$, we take

$$W_{i+1} := \begin{pmatrix} \tau_i I & 0 \\ 0 & \sigma_{i+1} I \end{pmatrix}, \quad \text{and} \quad Z_{i+1} := \begin{pmatrix} \phi_i I & 0 \\ 0 & \psi_{i+1} I \end{pmatrix}. \tag{3.2}$$

We suppose that ∂G_0 is *(strongly) monotone*, satisfying

$$\langle \partial G_0(x') - \partial G_0(x), x' - x \rangle \geq \gamma \|x' - x\|^2 \quad (x, x' \in X) \tag{G_0-SM}$$

for some $\gamma \geq 0$. Regarding J, we assume that ∇J exists and is *co-coercive* in the sense that for some $L \geq 0$ holds

$$\langle \nabla J(x') - \nabla J(x), x' - x \rangle \geq L^{-1} \|\nabla J(x') - \nabla J(x)\|^2 \quad (x, x' \in X). \tag{J-CO}$$

(We allow $L = 0$ for the case $J = 0$.)

We also introduce

$$\Xi_{i+1}(\gamma) := \begin{pmatrix} 2\tau_i\gamma I & 2\tau_i K^* \\ -2\sigma_{i+1}K & 0 \end{pmatrix}, \quad \text{and} \quad Q_{i+1}(L) := \begin{pmatrix} L\tau_i I & 0 \\ 0 & 0 \end{pmatrix},$$

which are operator measures of strong monotonicity and smoothness of H. Finally, we introduce the forward–step preconditioner with respect to J, familiar from Example 2.5 as

$$V^J_{i+1}(u) := \begin{pmatrix} \tau_i(\nabla J(x^i) - \nabla J(x)) \\ 0 \end{pmatrix}. \tag{3.3}$$

Theorem 3.1. *Let us be given $K \in \mathcal{L}(X;Y)$, $G = G_0 + J \in \Gamma(X)$, and $F^* \in \Gamma(Y)$ on Hilbert spaces X and Y. Suppose G_0 satisfies (G_0-SM) for some $\gamma \geq 0$, and J satisfies (J-CO) for some $L \geq 0$. For each $i \in \mathbb{N}$, let $\tau_i, \phi_i, \sigma_{i+1}, \psi_{i+1} > 0$. Also take $V'_{i+1} : X \times Y \to X \times Y$, and $M_{i+1} \in \mathcal{L}(X\times Y; X\times Y)$. Let H given by (3.1), Z_{i+1} and W_{i+1} by (3.2), and V_{i+1} by (2.2). Suppose (PP, p.11) is solvable, and denote the iterates by $u^i = (x^i, y^i)$. Then (CI, p.12) and (DI, p.12) hold if $Z_{i+1}M_{i+1}$ is self-adjoint, and*

$$\text{(CI-}\Gamma\text{)} \quad \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}(M_{i+1} - Q_{i+1}(L/2))} + \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}(\Xi_{i+1}(\gamma)+M_{i+1})-Z_{i+2}M_{i+2}} + \langle V'_{i+1}(u^{i+1}) - V^J_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}} \geq -\Delta_{i+1}(\widehat{u}).$$

Proof. Using (J-CO), similarly to (2.9) we derive

$$\langle \nabla J(x^i) - \nabla J(\widehat{x}), x^{i+1} - \widehat{x}\rangle \geq -\frac{L}{4}\|x^{i+1} - x^i\|^2$$

Using (3.3), therefore

$$\langle V^J_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}} \geq -\frac{L\phi_i\tau_i}{4}\|x^{i+1} - x^i\|^2 - \phi_i\tau_i\langle \nabla J(x^{i+1}) - \nabla J(\widehat{x}), x^{i+1} - \widehat{x}\rangle.$$

With this, the monotonicity of ∂F^*, and (G_0-SM), we observe (CI-Γ) to imply

$$\frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{Z_{i+1}(\Xi_{i+1}(0)+M_{i+1})-Z_{i+2}M_{i+2}} + \langle \partial G(x^{i+1}) - \partial G(\widehat{x}), x^{i+1} - \widehat{x}\rangle_{\Phi_i T_i} + \langle \partial F^*(y^{i+1}) - \partial F^*(\widehat{y}), y^{i+1} - \widehat{y}\rangle_{\Psi_{i+1}\Sigma_{i+1}} + \langle V'_{i+1}(u^{i+1}), u^{i+1} - \widehat{u}\rangle_{Z_{i+1}} \geq -\Delta_{i+1}(\widehat{u}). \tag{3.4}$$

Here pay attention to the fact that (3.4) employs $\Xi_{i+1}(0)$ while (CI-Γ) employs $\Xi_{i+1}(\gamma)$. If we show that (CI, p.12) follows from (3.4), then (DI, p.12) follows from Theorem 2.1. Indeed, using the expansion

$$Z_{i+1}W_{i+1} = \begin{pmatrix} \phi_i\tau_i I & 0 \\ 0 & \psi_{i+1}\sigma_{i+1}I \end{pmatrix},$$

we expand for any $\widetilde{u} = (\widetilde{x}, \widetilde{y})$ that

$$\begin{aligned}\langle Z_{i+1}W_{i+1}(H(u^{i+1}) - H(\widetilde{u})), u^{i+1} - \widetilde{u}\rangle \\ = \phi_i\tau_i\langle \partial G(x^{i+1}) - \partial G(\widetilde{x}), x^{i+1} - \widetilde{x}\rangle + \psi_{i+1}\sigma_{i+1}\langle \partial F^*(y^{i+1}) - \partial F^*(\widetilde{y}), y^{i+1} - \widetilde{y}\rangle \\ + \phi_i\tau_i\langle K^*(y^{i+1} - \widetilde{y}), x^{i+1} - \widetilde{x}\rangle - \psi_{i+1}\sigma_{i+1}\langle K(x^{i+1} - \widetilde{x}), y^{i+1} - \widetilde{y}\rangle.\end{aligned}$$

With the help of $\Xi_{i+1}(0)$ we then obtain

$$\begin{aligned}\langle H(u^{i+1}) - H(\widetilde{u}), u^{i+1} - \widetilde{u}\rangle_{Z_{i+1}W_{i+1}} \geq \frac{1}{2}\|u^{i+1} - \widetilde{u}\|_{Z_{i+1}\Xi_{i+1}(0)} \\ + \phi_i\tau_i\langle \partial G(x^{i+1}) - \partial G(\widetilde{x}), x^{i+1} - \widetilde{x}\rangle + \psi_{i+1}\sigma_{i+1}\langle \partial F^*(y^{i+1}) - \partial F^*(\widetilde{y}), y^{i+1} - \widetilde{y}\rangle.\end{aligned}$$

Inserting this into (3.4), we obtain (CI, p.12). Then we apply Theorem 2.1. □

Remark 3.2. *For gap estimates and other extensions, we refer to [31].*

3.2 EXAMPLES OF PRIMAL–DUAL METHODS

We now look at several known methods for the saddle point problem (1.7).

Example 3.1 (The primal–dual method of Chambolle and Pock [5]). This method consists of iterating the system

$$x^{i+1} := (I + \tau_i\partial G)^{-1}(x^i - \tau_i K^* y^i), \tag{3.5a}$$

$$\bar{x}^{i+1} := \omega_i(x^{i+1} - x^i) + x^{i+1}, \tag{3.5b}$$

$$y^{i+1} := (I + \sigma_{i+1}\partial F^*)^{-1}(y^i + \sigma_{i+1}K\bar{x}^{i+1}). \tag{3.5c}$$

In the basic version of the algorithm, $\omega_i = 1$, $\tau_i \equiv \tau_0 > 0$, and $\sigma_i \equiv \sigma_0 > 0$, assuming the step length parameters to satisfy

$$\tau_0\sigma_0\|K\|^2 < 1. \tag{3.6}$$

The iterates convergence weakly, and the method has $O(1/N)$ rate for ann ergodic duality gap, which we skip in these notes; details can be found in [5, 31]. If G is strongly convex with factor γ, we may accelerate

$$\omega_i := 1/\sqrt{1 + 2\gamma\tau_i}, \quad \tau_{i+1} := \tau_i\omega_i, \quad \text{and} \quad \sigma_{i+1} := \sigma_i/\omega_i. \tag{3.7}$$

This yields $O(1/N^2)$ convergence of $\|x^N - \widehat{x}\|^2$ to zero.

Verification. We formulate the method in our proximal point framework with $J = 0$ and $G_0 = G$ following [15, 32] by taking as the preconditioner

$$M_{i+1} = \begin{pmatrix} I & -\tau_i K^* \\ -\sigma_i K & I \end{pmatrix} \quad \text{and} \quad V'_{i+1} = 0.$$

Taking $\Delta_{i+1}(\widehat{u}) := -\frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}$, we reduce (CI-Γ) to

$$\frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{D_{i+2}} \geq 0 \quad \text{for} \quad D_{i+2} := Z_{i+1}(\Xi_{i+1}(\gamma) + M_{i+1}) - Z_{i+2}M_{i+2}. \tag{3.8}$$

We may expand

$$Z_{i+1}M_{i+1} = \begin{pmatrix} \phi_i I & -\phi_i\tau_i K^* \\ -\psi_{i+1}\sigma_i K & \psi_{i+1} I \end{pmatrix}, \quad \text{and} \tag{3.9a}$$

$$D_{i+2} = \begin{pmatrix} (\phi_i(1 + 2\widetilde{\gamma}\tau_i) - \phi_{i+1})I & (\phi_i\tau_i + \phi_{i+1}\tau_{i+1})K^* \\ (\psi_{i+2}\sigma_{i+1} - 2\psi_{i+1}\sigma_{i+1} - \psi_{i+1}\sigma_i)K & (\psi_{i+1} - \psi_{i+2})I \end{pmatrix}. \tag{3.9b}$$

We have $\|\cdot\|_{D_{i+2}} = 0$ (but not $D_{i+2} = 0$, as the former depends on the off-diagonals cancelling out), and $Z_{i+1}M_{i+1}$ is self-adjoint, if for some constant ψ we take

$$\phi_{i+1} := \phi_i(1 + 2\widetilde{\gamma}\tau_i), \quad \tau_i := \phi_i^{-1/2}, \quad \sigma_i := \phi_i\tau_i/\psi, \quad \text{and} \quad \psi_{i+1} := \psi. \tag{3.10}$$

This gives the acceleration scheme (3.7). Moreover, for any $\delta \in (0, 1)$ holds

$$Z_{i+1}M_{i+1} \geq \begin{pmatrix} \delta\phi_i I & 0 \\ 0 & \psi I - (1-\delta)^{-1}KK^* \end{pmatrix}. \tag{3.11}$$

Thus $Z_{i+1}M_{i+1} \geq 0$ if $\psi \geq (1-\delta)^{-1}\|K\|^2$. By (3.10), $\sigma_i\tau_i = 1/\psi$. Since this fixes the ratio of σ_i to τ_i, we need to take $\psi := 1/(\sigma_0\tau_0)$ as well as $\delta := 1 - \sigma_0\tau_0\|K\|^2$. Through the positivity of δ, we recover the initialisation condition (3.6).

Theorem 3.1 and Corollary 2.6 show weak convergence of the iterates without a rate. If G is strongly convex with factor $\gamma \geq 0$, so that also $\widetilde{\gamma} > 0$, the results in [5, 32] show that τ_N is of the order $O(1/N)$, and consequently ϕ_N is of the order $\Theta(N^2)$. By Corollary 2.3, $\|x^N - \widehat{x}\|^2$ converges to zero at the rate $O(1/N^2)$. □

Example 3.2 (Chambolle–Pock with a forward step). Suppose $G = G_0 + J$ with G (strongly) convex with factor $\gamma \geq 0$, and ∇J Lipschitz with factor L. In [6], the Chambolle–Pock method was extended to take forward steps with respect to J. With everything else as in Example 3.1, take

$$V'_{i+1}(u) := V^J_{i+1}(u) = (\tau_i(\nabla J(x^i) - \nabla J(x)), 0).$$

Then (PP, p.11) can be rearranged as

$$x^{i+1} := (I + \tau_i\partial G_0)^{-1}(x^i - \tau_i\nabla J(x^i) - \tau_i K^* y^i), \tag{3.12}$$

$$\bar{x}^{i+1} := \omega_i(x^{i+1} - x^i) + x^{i+1}, \tag{3.13}$$

$$y^{i+1} := (I + \sigma_{i+1}\partial F^*)^{-1}(y^i + \sigma_{i+1}K\bar{x}^{i+1}). \tag{3.14}$$

The method inherits the convergences properties of Example 3.1 if we use the step

length update rules (3.7), and initialise $\tau_0, \sigma_0 > 0$ subject to (3.6), and

$$0 < \theta := 1 - L\tau_0/(1 - \tau_0\sigma_0\|K\|^2). \tag{3.15}$$

Verification. With D_{i+2} as in (3.8), the condition (CI-Γ, p.18) becomes

$$\frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} - \frac{\tau_i\phi_i L}{4}\|x^{i+1} - x^i\|^2 + \frac{1}{2}\|u^{i+1} - \widehat{u}\|^2_{D_{i+2}} \geq -\Delta_{i+1}(\widehat{u}). \tag{3.16}$$

The rules (3.10) force $\|\cdot\|_{D_{i+2}} = 0$. We take $\Delta_{i+1}(\widehat{u}) = -\frac{\theta}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}}$ for some $\theta > 0$, and deduce using Cauchy's inequality that (3.16) holds if

$$(1-\theta)Z_{i+1}M_{i+1} \geq \tau_i\phi_i L\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}.$$

Recalling (3.11), this is true if $(1-\theta)\delta\phi_i \geq \tau_i\phi_i L$ and $\psi \geq (1-\delta)^{-1}\phi_i\tau_i^2\|K\|^2$. Further recalling (3.10), and observing that $\{\tau_i\}$ is non-increasing, we only have to satisfy $(1-\theta)(1-\tau_0\sigma_0\|K\|^2) \geq L\tau_0$. Otherwise put, we obtain (3.15). □

Example 3.3 (Alternating Directions Method of Multipliers, briefly). The classical ADMM [13] and Douglas–Rachford splitting [12] are known to be related to the Chambolle–Pock method; in fact the Chambolle–Pock method is a preconditioned ADMM [5]. From [3, Section 5], we can deduce that compared to the Chambolle–Pock method, the ADMM merely has the sign of K reversed in

$$M_{i+1} = \begin{pmatrix} I & \tau_i K \\ \sigma_i K & I \end{pmatrix}.$$

Taking $\tau_i = \tau_0$ and $\sigma_i = \sigma_0$ constant and satisfying (3.6), the iterates converge weakly. Acceleration can provide $O(1/N)$ convergence of $\|x^N - \widehat{x}\|^2$.

Verification. Following Example 3.1, we now expand

$$D_{i+2} = \begin{pmatrix} (\phi_i(1+2\widetilde{\gamma}\tau_i) - \phi_{i+1})I & (3\phi_i\tau_i - \phi_{i+1}\tau_{i+1})K^* \\ (\psi_{i+1}\sigma_i - 2\psi_{i+1}\sigma_{i+1} - \psi_{i+2}\sigma_{i+1})K & (\psi_{i+1} - \psi_{i+2})I \end{pmatrix}.$$

This time $\|\cdot\|_{D_{i+2}} = 0$ and $Z_{i+1}M_{i+1}$ is self-adjoint if we take

$$\phi_{i+1} := \phi_i(1+2\widetilde{\gamma}\tau_i), \quad \tau_{i+1} := \tau_i\phi_i/\phi_{i+1}, \quad \sigma_i := \phi_i\tau_i/\psi, \quad \text{and} \quad \psi_{i+1} := \psi. \tag{3.17}$$

If $\widetilde{\gamma} = 0$, which corresponds to the standard ADMM with fixed step lengths, it is easy to retrace the steps of Example 3.1 to prove weak convergence (without a rate). If $\widetilde{\gamma} \neq 0$, we obtain $\phi_{N+1} = \phi_N + 2\widetilde{\gamma}\tau_{N-1}\phi_{N-1} = \phi_N + 2\widetilde{\gamma}\tau_0\phi_0 = \phi_0 + 2N\widetilde{\gamma}\tau_0\phi_0$. Therefore, the acceleration scheme (3.17) only gives the rate $O(1/N)$. □

Example 3.4 (GIST). Suppose $G(x) = \frac{1}{2}\|f - Ax\|^2$, $\|A\| < \sqrt{2}$, and $\|K\| \leq 1$. Take

$$V'_{i+1}(u) := \begin{pmatrix} \nabla G(x^i) - \nabla G(x) \\ 0 \end{pmatrix}, \quad \text{and} \quad M_{i+1} := \begin{pmatrix} I & 0 \\ 0 & I - KK^* \end{pmatrix}.$$

With $T_i := I$ and $\Sigma_{i+1} := I$, we then obtain the Generalised Iterative Soft Thresholding (GIST) algorithm of [19]

$$\begin{aligned} y^{i+1} &:= (I + \partial F^*)^{-1}((I - KK^*)y^i + K(x^i - \nabla G(x^i))), \\ x^{i+1} &:= x^i - \nabla G(x^i) - K^* y^{i+1}. \end{aligned}$$

The iterates $\{x^i\}_{i\in\mathbb{N}}$ converge weakly to $\widehat{x}$.

Exercise 3.1. *Using Theorem 3.1 and Corollary 2.6, prove the convergence of GIST.*

3.3 NON-LINEAR FORWARD OPERATORS

Suppose we want to solve a problem of the form

$$\min_x \ \frac{1}{2}\|z - T(x)\|^2 + F(Ax),$$

where $T \in C^1(X; Y)$ is non-linear. Recalling Example 1.3, we can convert this problem into the form

$$\min_x \max_{y,\lambda} \ \langle (Ax, T(x)), (y, \lambda)\rangle - \left[\frac{1}{2}\|\lambda\|^2 + \langle z, \lambda\rangle + F^*(y)\right].$$

Therefore we have a problem of the form (1.7) with K non-linear, that is

$$\min_x \max_y \ G(x) + \langle K(x), y\rangle - F^*(y). \tag{3.18}$$

It follows from properties of convex conjugates and, e.g., [7, Theorem 2.3.10 and Proposition 2.3.6] applied to $F(K(x))$ that the first-order necessary optimality conditions for this problem can be written

$$-[\nabla K(\widehat{x})]^* \widehat{y} \in \partial G(\widehat{x}), \quad K(\widehat{x}) \in F^*(\widehat{y}). \tag{3.19}$$

In the linear case we had $-K^*\widehat{y} \in \partial G(\widehat{x})$ and $K\widehat{x} \in F^*(\widehat{y})$. Making corresponding changes to the PDHGM from (1.9), we are led to the NL-PDHGM algorithm [26]

$$x^{i+1} := (I + \tau_i \partial G)^{-1}(x^i - \tau_i [\nabla K(x^i)]^* y^i), \tag{3.20a}$$

$$\bar{x}^{i+1} := \omega_i (x^{i+1} - x^i) + x^{i+1}, \tag{3.20b}$$

$$y^{i+1} := (I + \sigma_{i+1} \partial F^*)^{-1}(y^i + \sigma_{i+1} K(\bar{x}^{i+1})). \tag{3.20c}$$

Corresponding to the optimality conditions (3.19), we define the set-valued operator $H : X \times Y \rightrightarrows X \times Y$ for $u = (x, y)$ as

$$H(u) := \begin{pmatrix} \partial G(x) + [\nabla K(x)]^* y \\ \partial F^*(y) - K(x) \end{pmatrix}, \tag{3.21}$$

such that $0 \in H(\widehat{u})$ encodes first-order necessary optimality conditions for our problem.

As in Example 3.1, we define for some $\phi_i, \psi_{i+1} > 0$ the step length operator and testing operators

$$W_{i+1} := \begin{pmatrix} \tau_i I & 0 \\ 0 & \sigma_{i+1} I \end{pmatrix}, \quad \text{and} \quad Z_{i+1} := \begin{pmatrix} \phi_i I & 0 \\ 0 & \psi_{i+1} I \end{pmatrix}.$$

We also define the non-linear preconditioner $V_{i+1}(u) := V'_{i+1}(u) + M_{i+1}(u - u^i)$ by

$$V'_{i+1}(u) := W_{i+1} \begin{pmatrix} [\nabla K(x^i) - \nabla K(x)]^* y \\ K(x) - K([\![x, x^i]\!]^\omega) - \nabla K(x^i)(x - [\![x, x^i]\!]^{\omega_i}) \end{pmatrix}, \tag{3.22}$$

and

$$M_{i+1} := \begin{pmatrix} I & -\tau_i [\nabla K(x^i)]^* \\ -\omega_i \sigma_{i+1} \nabla K(x^i) & I \end{pmatrix}, \tag{3.23}$$

where $\omega_i := \psi_{i+1}^{-1} \phi_i \tau_i$, and $\bar{x}^{i+1} := [\![x^{i+1}, x^i]\!]^{\omega_i}$ for $[\![x, x^i]\!]^\omega := x + \omega(x - x^i)$.

Now (3.20) can be written in the standard form

$$0 \in W_{i+1} H(u^{i+1}) + V_{i+1}(u^{i+1}). \tag{3.24}$$

In finite dimensions, the convergence of this method is proved in [26], on the assumption of *metric regularity* of H. For the infite-dimensional case, see [9, 10]. More recent results in [8] are based on the theory presented here.

3.4 SPATIAL ADAPTATION AND STOCHASTIC METHODS

Let us suppose G and F^* are separable as

$$G(x) = \sum_{j=1}^{m} G_j(P_j x), \quad \text{and} \quad F^*(y) = \sum_{\ell=1}^{n} F_\ell^*(Q_\ell y).$$

for some projection oeprators $P_1, \dots, P_m$ in X with $\sum_{j=1}^m P_j = I$ and $P_j P_i = 0$ if $i \neq j$. Likewise, $Q_1, \dots, Q_n$ are similarly projection operators in Y. Similarly to G and F^*, we assume all the component functions G_j and F_ℓ^* to be convex.

Let us replace the step length operators τ_i and σ_{i+1} by blockwise operators

$$T_i := \sum_{j \in S(i)} \tau_{j,i} P_j, \quad \text{and} \quad \Sigma_{i+1} := \sum_{\ell \in V(i+1)} \sigma_{\ell,i+1} Q_\ell, \quad (i \geq 0),$$

where $\tau_{j,i}, \sigma_{\ell,i+1} \geq 0$ and $S(i) \subset \{1, \dots, m\}$, $V(i+1) \subset \{1, \dots, n\}$. These subsets we allow to be random.

Then we can derive stochastic and "spatiall adaptive" algorithms that update each of the "blocks" $P_j x$ and $Q_\ell x$ separately. In these algorithms, it will generally not be possible to satisfy

$$Z_{i+1}(M_{i+1} + \Xi_{i+1}) \geq Z_{i+2} M_{i+2}.$$

Therefore the penalty Δ_{i+1} in Theorem 3.1 is non-zero, and produces "mixed" $O(1/N + 1/N^2)$ convergence rates for problems where only some of the functions G_j are strongly convex. For details we refer to [28].

4 FASTER CONVERGENCE FROM REGULARITY

[In these notes, this last chapter, extracted from [30], is done at a somewhat more abstract level than in the lectures. A more introductory version is a work in progress.]

What is the weakest useful form of regularity of a set-valued map H? In particular, if $0 \in H(\widehat{u})$ for $\widehat{u} = (\widehat{x}, \widehat{y})$ encodes optimality conditions of a saddle point problem (1.7), what regularity property is useful for showing faster—improved—convergence of optimisation methods, compared to that obtainable by the basic analysis of the previous chapters?

A starting point for the regularity of set-valued maps is to extend the Lipschitz property of single-valued maps. One such approach is the *Aubin*, *pseudo-Lipschitz*, or *Lipschitz-like* property. When we are interested in the stability of the optimality condition $0 \in H(\widehat{u})$, it is typically more beneficial to study the Aubin property of the inverse H^{-1}. This is called the *metric regularity* of H at (or near) a point $(\widehat{u}, \widehat{w}) \in \operatorname{Graph} H$. In this property, both u and w are allowed to vary in the criterion

$$\kappa \operatorname{dist}(w, H(u)) \geq \operatorname{dist}(u, H^{-1}(w)) \quad (u \in \mathcal{U},\ w \in \mathcal{W}),$$

which is assumed to hold for some $\kappa > 0$, and neighbourhoods $\mathcal{U} \ni \widehat{u}$ and $\mathcal{W} \ni \widehat{w}$. Metric regularity is equivalent to *openness at a linear rate* near $(\widehat{u}, \widehat{w})$, and holds for smooth maps by the class Lyusternik–Graves theorem [see, e.g., 17]. It is too strong a property to be satisfied in many applications. In [26], we used metric regularity to show the convergence of the NL-PDHGM from [26]. In a newer analysis [8] we no longer need it, instead using more direct monotonicity-based analysis. In this chapter, we will also look at a weaker notion of *(partial) submonotonicity*.

These notions are motivated by *metric subregularity*. It allows much more leeway for H by fixing $w = \widehat{w}$. In other words, we require

$$\kappa \operatorname{dist}(\widehat{w}, H(u)) \geq \operatorname{dist}(u, H^{-1}(\widehat{w})) \quad (u \in \mathcal{U}). \tag{4.1}$$

The counterpart of metric subregularity that relaxes the Aubin property is known as *calmness* or the *upper Lipschitz* property [23]. We refer to the books [1, 11, 17, 20, 25] for further information on these and other related properties. These include the Mordukhovich criterion that allows verifying the Aubin property or metric regularity through coderivative considerations.

Before introducing our notions of regularity, we improve Theorem 2.1 to study convergence to the entire set $H^{-1}(0)$ instead of a specific point $\widehat{u}$. This is as we might expect from (4.1).

4.1 CONVERGENCE TO SET OF CRITICAL POINTS

We continue with the abstract setup of Chapter 2. Specifically, we want to solve (PP, p.11), which is

$$0 \in W_{i+1}H(u^{i+1}) + V'_{i+1}(u^{i+1}) + M_{i+1}(u^{i+1} - u^i).$$

The next result modifies Theorem 2.1 to replace $\|u - \widehat{u}\|^2_{Z_{i+1}M_{i+1}}$ for a fixed $\widehat{u} \in H^{-1}(0)$ by the distance $\operatorname{dist}^2_{Z_{i+1}M_{i+1}}(u; H^{-1}(0))$ to the solution set.

Theorem 4.1. *On a Hilbert space U, let $H : U \rightrightarrows U$, and $W_{i+1}, M_{i+1}, Z_{i+1} \in \mathcal{L}(U;U)$, as well as $V'_{i+1} : U \to U$ for $i \in \mathbb{N}$. Suppose (PP, p.11) is solvable for V_{i+1} as in (2.2), and denote the iterates by $\{u^i\}_{i\in\mathbb{N}}$. If $Z_{i+1}M_{i+1}$ is self-adjoint, and*

$$\begin{aligned}(\mathrm{CI}^*)\quad & \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} \\ & + \inf_{u^*\in H^{-1}(0)} \left(\frac{1}{2}\|u^{i+1} - u^*\|^2_{Z_{i+1}M_{i+1}} + \langle W_{i+1}H(u^{i+1}) + V'_{i+1}(u^{i+1}), u^{i+1} - u^*\rangle_{Z_{i+1}} \right) \\ & \qquad \geq \frac{1}{2}\operatorname{dist}^2_{Z_{i+2}M_{i+2}}(u^{i+1}; H^{-1}(0)) - \Delta_{i+1}\end{aligned}$$

for all $i \in \mathbb{N}$ and some $\Delta_{i+1} \in \mathbb{R}$, then

$$(\mathrm{DI}^*)\qquad \frac{1}{2}\operatorname{dist}^2_{Z_{N+1}M_{N+1}}(u^N, H^{-1}(0)) \leq \frac{1}{2}\operatorname{dist}^2_{Z_1M_1}(u^0, H^{-1}(0)) + \sum_{i=0}^{N-1}\Delta_{i+1} \quad (N \geq 1).$$

Proof. Let $u^* \in H^{-1}(0)$ be arbitrary. Inserting (PP, p.11) into (CI^*), we obtain

$$\begin{aligned}(4.2)\quad & \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} + \inf_{u^*\in H^{-1}(0)} \left(\frac{1}{2}\|u^{i+1} - u^*\|^2_{Z_{i+1}M_{i+1}} - \langle u^{i+1} - u^i, u^{i+1} - u^*\rangle_{Z_{i+1}M_{i+1}} \right) \\ & \qquad \geq \frac{1}{2}\operatorname{dist}^2_{Z_{i+2}M_{i+2}}(u^{i+1}; H^{-1}(0)) - \Delta_{i+1}.\end{aligned}$$

We recall for general self-adjoint M the three-point formula (2.4), that is

$$\langle u^{i+1} - u^i, u^{i+1} - u^*\rangle_M = \frac{1}{2}\|u^{i+1} - u^i\|^2_M - \frac{1}{2}\|u^i - u^*\|^2_M + \frac{1}{2}\|u^{i+1} - u^*\|^2_M.$$

Using this with $M = Z_{i+1}M_{i+1}$, we rewrite (4.2) as

$$\frac{1}{2}\operatorname{dist}^2_{Z_{i+1}M_{i+1}}(u^i; H^{-1}(0)) \geq \frac{1}{2}\operatorname{dist}^2_{Z_{i+2}M_{i+2}}(u^{i+1}; H^{-1}(0)) - \Delta_{i+1}.$$

Summing over $i = 0, \ldots, N-1$, we obtain the claim. □

It is possible to obtain weak convergence from this result [see 30], but we concentrate on strong results.

4.2 RATES FROM STRONG MONOTONICITY

Suppose for some $\Xi_{i+1} \in \mathcal{L}(U;U)$ we have strong monotonicity of the form

$$\langle H(u) - H(u'), u - u' \rangle_{Z_{i+1}W_{i+1}} \geq \|u - u'\|_{Z_{i+1}\Xi_{i+1}} \quad (u, u' \in U), \tag{4.3}$$

If also $V'_{i+1} = 0$, or if V'_{i+1} can otherwise be approximated away from (CI*), then we see that (CI*) holds with the penalty $\Delta_{i+1} = 0$ if we secure

$$Z_{i+1}(M_{i+1} + \Xi_{i+1}) \geq Z_{i+2}M_{i+2}. \tag{4.4}$$

From (DI*) we are then able to obtain convergence rates for $\operatorname{dist}^2(u^N; H^{-1}(0))$.

Example 4.1 ($O(1/N^2)$ convergence rate). Suppose $M_{i+1} = I$, $Z_{i+1} = \phi_{i+1}I$, and $\Xi_{i+1} = \gamma\phi_{i+1}^{-1/2}$ for some $\phi_{i+1} > 0$ and $\gamma > 0$. Then (4.4) as an equality gives the rule $\phi_{i+1} := \phi_i + \gamma\phi_{i+1}^{1/2}$. From this it is possible to show that $\phi_N \geq CN^2$ for some constant $C > 0$ [5, 32]. We therefore deduce from (DI*) the $O(1/N^2)$ convergence of $\operatorname{dist}^2(u^N; H^{-1}(0))$ to zero.

Example 4.2 (Linear convergence rate). Suppose $M_{i+1} = I$, $Z_{i+1} = \phi_{i+1}I$, and $\Xi_{i+1} = \gamma$ for some $\phi_{i+1} > 0$ and $\gamma > 0$. Then (4.4) as an equality gives the rule $\phi_{i+1} := \phi_i(1 + \gamma)$. Clearly then $\phi_N \geq \phi_0(1 + \gamma)^N$. In other words, we obtain from (DI*) linear convergence of $\operatorname{dist}^2(u^N; H^{-1}(0))$ to zero.

4.3 RATES FROM SUBMONOTONICITY

In (CI*) we can fix $u' = \widehat{u}$, so do not need the full power of monotonicity of the form (4.3). Indeed, we are led to thinking we can take the infimum over $u' \in H^{-1}(0)$. However, we have to be careful to keep this minimisation compatible with $\operatorname{dist}^2_{Z_{i+2}M_{i+2}}(u^{i+1}; H^{-1}(0))$. We therefore introduce the following concept.

Definition 4.1. Let $N, M, \Xi \in \mathcal{L}(U;U)$ with $M \geq 0$. We say that $T : U \rightrightarrows U$ is (Ξ, N, M)-*partially strongly submonotone at* $(\widehat{u}, \widehat{w}) \in \operatorname{Graph} T$ if there exists a neighbourhood $\mathcal{U} \ni \widehat{u}$ where for all $u \in \mathcal{U}$ and $w \in T(u)$ holds

$$\inf_{u^* \in T^{-1}(\widehat{w})} \left(\langle w - \widehat{w}, u - u^* \rangle_N + \|u - u^*\|^2_{M-\Xi} \right) \geq \operatorname{dist}^2_M(u, T^{-1}(\widehat{w})). \tag{PSM}$$

If $\Xi = M$, we say that T is (N, M)-*strongly submonotone*. If $\Xi = 0$, we say that T is (N, M)-*submonotone*.

Remark 4.2 (Submonotonicity from monotonicity). *(Ξ, M, N)-partial strong submonotonicity for any $M \geq 0$ is implied by*

$$\langle w - \widehat{w}, u - u^* \rangle_N \geq \|u - u^*\|_\Xi^2 \quad (u \in \mathcal{U},\ w \in T(u),\ u^* \in T^{-1}(\widehat{w})).$$

Remark 4.3 (Limited dependence on base point). *Submonotonicity only depends on $\widehat{u}$ through $\mathcal{U}$.*

Remark 4.4 (Scaling invariance). *For any factor $\alpha > 0$, (Ξ, N, M)-partial strong submonotonicity is equivalent to $(\alpha\Xi, \alpha N, \alpha M)$-partial strongly submonotonicity*

Returning to the preconditioned proximal point method (PP, p.11), the next result shows how $Z_{i+2}M_{i+2}$ can be made to grow based on partial strong submonotonicity, and therefore how this can help us obtain convergence rates.

Corollary 4.5. *On a Hilbert space U, let $H : U \rightrightarrows U$, $V'_{i+1} : U \to U$, and $M_{i+1}, W_{i+1}, Z_{i+1}, \Xi_{i+1} \in \mathcal{L}(U;U)$ with $Z_{i+1}M_{i+1} \geq 0$ self-adjoint for all $i \in \mathbb{N}$. Suppose (PP, p.11) is solvable for the iterates $\{u^i\}_{i\in\mathbb{N}}$. If H is $(Z_{i+1}\Xi_{i+1}, 2Z_{i+1}W_{i+1}, Z_{i+2}M_{i+2})$-partially strongly submonotone at some $(\widehat{u}, 0) \in \operatorname{Graph} H$, and*

$$\text{(CI-M)}\quad \frac{1}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1} - u^*\|^2_{Z_{i+1}(M_{i+1}+\Xi_{i+1})-Z_{i+2}M_{i+2}} + \langle V'_{i+1}(u^{i+1}), u^{i+1} - u^* \rangle_{Z_{i+1}} \geq -\Delta_{i+1}$$

for some $\Delta_{i+1} \in \mathbb{R}$ for all $i \in \mathbb{N}$ and $u^* \in H^{-1}(0)$, then (DI, p.26) holds provided $\{u^i\}_{i=0}^N \subset \mathcal{U}$ for the neighbourhood $\mathcal{U}$ of partial strong submonotonicity.*

Proof. By the assumed partial strong submonotonicity, for all $u^* \in H^{-1}(0)$ holds

$$\langle H(u^{i+1}), u^{i+1} - u^* \rangle_{Z_{i+1}W_{i+1}} + \frac{1}{2}\|u^{i+1} - u^*\|_{Z_{i+2}M_{i+2}-Z_{i+1}\Xi_{i+1}} \geq \frac{1}{2}\operatorname{dist}^2_{Z_{i+2}M_{i+2}}(u^{i+1}; H^{-1}(0)).$$

Summing this with (CI-M), and taking the infimum over $u^* \in H^{-1}(0)$, we obtain (CI*, p.26). Then we just apply Theorem 4.1. □

Example 4.3 (Basic proximal point method, submonotonicity). Suppose for some $\tau > 0$ and $\xi \geq 0$ that H is $(\xi I, 2\tau I, (1+\xi)I)$-partially strongly submonotone at $(\widehat{u}, 0) \in \operatorname{Graph} H$. This is to say

$$\inf_{u^* \in H^{-1}(0)} \left(2\tau\langle w, u - u^* \rangle + \|u - u^*\|^2\right) \geq (1+\xi)\operatorname{dist}^2(u, H^{-1}(0)) \quad (u \in \mathcal{U},\ w \in H(u)).$$

Take $M_{i+1} := I$, $V'_{i+1} = 0$, as well as $W_{i+1} := \tau I$. Then (PP, p.11) describes the basic proximal point method $0 \in H(u^{i+1}) + \tau^{-1}(u^{i+1} - u^i)$. Its iterates satisfy $\operatorname{dist}^2(u^N; H^{-1}(0)) \leq (1+\xi)^{-N}\operatorname{dist}^2(u^0; H^{-1}(0))$.

Verification. We take $\Xi_{i+1} := \xi I$. $Z_{i+1} := \phi_i I$ for some $\phi_i > 0$. Then $Z_{i+1}(M_{i+1} + \Xi_{i+1}) = Z_{i+2}M_{i+2}$ if we update $\phi_{i+1} := \phi_i(1+\xi)$ for some $\phi_0 > 0$. By Remark 4.4, $(\xi I, 2\tau I, (1+\xi)I)$-partial strong submonotonicity is equivalent to $(\phi_i\xi I, 2\phi_i\tau I, \phi_i(1+\xi)I)$-partial strong submonotonicity. Since $\phi_{i+1} := \phi_i(1+\xi)$, by our definitions of Z_{i+1}, M_{i+1}, and W_{i+1}, we obtain the required $(Z_{i+1}\Xi_{i+1}, 2Z_{i+1}W_{i+1}, Z_{i+2}M_{i+2})$-partial strong submonotonicity. Consequently (CI-M) holds with $\Delta_{i+1} \equiv 0$. This is to say $\operatorname{dist}^2(u^N; H^{-1}(0)) \leq (\phi_0/\phi_N)\operatorname{dist}^2(u^0; H^{-1}(0))$. Now we use $\phi_{i+1} = \phi_i(1+\xi)$. □

Exercise 4.1 (Forward–backward splitting). *Suppose $H = H_0 + \nabla J$ with $J \in \Gamma(U)$ also L-smooth [see, e.g., 2]. With everything else as in Example 4.3, take $V'_{i+1}(u) := \tau(\nabla J(u^i) - \nabla J(u))$. Then (PP, p.11) describes the forward–backward splitting $u^{i+1} := (I + \tau H_0)^{-1}(u^i - \tau\nabla J(u^i))$, as we have already seen in Example 2.6. Show that as long as $L\tau \leq 2$, the convergence results of Example 4.3 apply.*

SOME FUNDAMENTAL CONVEX FUNCTIONS

Here we show on $\mathbb{R}$ that the subdifferentials of the indicator of the unit ball, and of the absolute value function are strongly submonotone. None of these subdifferentials are strongly monotone in the conventional sense. Throughout, with $(x^*, q^*) \in \operatorname{Graph}\partial G$, we consider $(I, \gamma I)$-strong submonotonicity, equivalently $(\gamma^{-1}I, I)$-strong submonotonicity for which we need to prove for some $\gamma > 0$ and neighbourhood $\mathcal{U}$ that

$$\langle \partial G(x) - q^*, x - x^*\rangle \geq \gamma \operatorname{dist}^2(x; [\partial G]^{-1}(q^*)) \quad (x \in \mathcal{U}). \tag{4.5}$$

We recall that $(I, \gamma I)$-strong submonotonicity implies $(\gamma I, I, I)$-partial strong submonotonicity.

Lemma 4.6. *Consider $G := \delta_{\operatorname{cl}\mathbb{B}(0,\alpha)}$, and let $(x^*, q^*) \in \operatorname{Graph}\partial G$. Then ∂G is $(I, \gamma I)$-strongly submonotone with*

$$\mathcal{U} := \operatorname{dom} G, \quad \text{and} \quad \gamma := \begin{cases} \|q^*\|/(2\alpha), & q^* \neq 0, \\ \infty, & q^* = 0. \end{cases}$$

Proof. Since $\langle \partial G(x), x - x^*\rangle \geq 0$, from (4.5) it suffices to prove for $x \in \operatorname{cl}\mathbb{B}(0,\alpha) = \operatorname{dom} G$ that

$$\langle q^*, x^* - x\rangle \geq \inf_{\widehat{x} \in [\partial G]^{-1}(q^*)} \gamma\|x - \widehat{x}\|^2. \tag{4.6}$$

If $q^* = 0$, then $[\partial G]^{-1}(q^*) = \operatorname{cl}\mathbb{B}(0,\alpha)$, so (4.6) trivially holds by the monotonicity of ∂G as a convex subdifferential [24].

Otherwise, if $q^* \neq 0$, necessarily $q^* = \beta x^*$ for some $\beta > 0$, and $\|x^*\| = \alpha$. Moreover, $[\partial G]^{-1}(q^*) = \{x^*\}$. Therefore (4.6) reads $\beta\langle x^*, x^* - x\rangle \geq \gamma\|x - x^*\|^2$. In other words $(\beta-\gamma)\|x^*\|^2 \geq \gamma\|x\|^2 + (\beta - 2\gamma)\langle x^*, x\rangle$. Since $\|x\| \leq \alpha$ and $\|x^*\| = \alpha$, this holds for $\beta \geq 2\gamma$. Since $q^* = \beta x^*$ and $\|x^*\| = \alpha$, this gives the maximal choice $\gamma = \|q^*\|/(2\alpha)$. □

Lemma 4.7. *Consider $G := |\cdot|$, and let $(x^*, q^*) \in \operatorname{Graph} \partial G$. Then ∂G is $(I, \gamma I)$-strongly submonotone for any $\gamma > 0$ in the neighbourhood $\mathcal{U} := ([1,-1] - q^*)/\gamma$.*

Proof. We need to prove (4.5). Since $\operatorname{dom} \partial G = [-1,1]$, it suffices to consider $x \in [-1,1]$. Clearly also $|q^*| \leq q$.

Consider first $q^* \in \{-1,1\}$. Now $[\partial G]^{-1}(q^*) = [-1,1]$, so (4.5) reduces to $\langle \partial G(x) - q^*, x - x^*\rangle$. This holds by the monotonicity of convex subdifferentials.

Consider then $|q^*| < 1$. Then $[\partial G]^{-1}(q^*) = \{x^*\} = \{0\}$. Then (4.5) holds if

$$\langle \partial G(x) - q^*, x\rangle \geq \gamma x^2. \tag{4.7}$$

If $x = 0$, this is clear. If $x > 1$, $\partial G(x) = \{1\}$, so (4.7) holds if $1 - q^* \geq \gamma x$. This holds if $x \leq (1-q^*)/\gamma$. Similarly, if $x < 1$, we obtain for (4.7) condition $-1 - q^* \leq \gamma x$. This holds when $x \geq (-1-q^*)/\gamma$. The conditions $x \leq (1-q^*)/\gamma$ and $-1-q^* \leq \gamma x$ give the expression for $\mathcal{U}$ in the statement of the lemma. □

4.4 RATES FROM ERROR BOUNDS

We now study an approach alternative to submonotonicity: the error bounds that we discussed in the introduction. Their essence is to prove for some $\kappa > 0$ that

$$\kappa\|u^{i+1} - u^i\| \geq \|u^{i+1} - \widehat{u}\|. \tag{EB}$$

One can see how this would improve (CI*, p.26) by allowing $Z_{i+1}M_{i+2}$ to grow faster. However, we generally cannot fix $\widehat{u}$, so would take the infimum over $\widehat{u} \in H^{-1}(0)$ above. In our case we also have to observe the changing metrics, and instead assume for some $\delta \in [0,1]$ and $P_{i+1} \in \mathcal{L}(U;U)$ with $Z_{i+2}M_{i+2} \geq Z_{i+1}P_{i+1}$ the *partial error bound*

$$\delta\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} + \operatorname{dist}^2_{Z_{i+2}M_{i+2} - Z_{i+1}P_{i+1}}(u^{i+1}, H^{-1}(0)) \geq \operatorname{dist}^2_{Z_{i+2}M_{i+2}}(u^{i+1}, H^{-1}(0)). \tag{PEB}$$

Corollary 4.8. *On a Hilbert space U, let $H : U \rightrightarrows U$, $V'_{i+1} : U \to U$, and $M_{i+1}, W_{i+1}, Z_{i+1}, \Xi_{i+1} \in \mathcal{L}(U;U)$ with $Z_{i+1}M_{i+1} \geq 0$ self-adjoint for all $i \in \mathbb{N}$. Suppose (PP, p.11) is solvable for the iterates $\{u^i\}_{i\in\mathbb{N}}$. If (PEB) holds, and*

$$\begin{aligned}\frac{1-\delta}{2}\|u^{i+1} - u^i\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1} - u^*\|^2_{Z_{i+1}(M_{i+1}+P_{i+1}) - Z_{i+2}M_{i+2}} \\ + \langle W_{i+1}H_{i+1}(u^{i+1}) + V'_{i+1}(u^{i+1}), u^{i+1} - u^*\rangle_{Z_{i+1}} \geq -\Delta_{i+1}(u^*)\end{aligned} \tag{CI-PEB}$$

for some $\Delta_{i+1} \in \mathbb{R}$ for all $i \in \mathbb{N}$ and $u^* \in H^{-1}(0)$, then (DI, p.26) holds.*

Proof. By (CI-PEB) for all $u^* \in H^{-1}(0)$ holds

$$\frac{\delta}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}} + \frac{1}{2}\|u^{i+1}-u^*\|_{Z_{i+2}M_{i+2}-Z_{i+1}P_{i+1}} \geq \frac{1}{2}\operatorname{dist}^2_{Z_{i+2}M_{i+2}}(u^{i+1}; H^{-1}(0)).$$

Summing this with (CI-PEB), and taking the infimum over $u^* \in H^{-1}(0)$, we obtain (CI*, p.26). Then we just apply Theorem 4.1. □

4.5 ERROR BOUNDS FROM METRIC SUBREGULARITY

An essential ingredient in proving the basic error bound (EB) is the *metric subregularity of H at $\widehat{u}$ for $\widehat{w} = 0$*: the existence of a neighbourhood $\mathcal{U} \ni \widehat{u}$ and $\kappa > 0$ such that

$$\kappa \operatorname{dist}(\widehat{w}, H(u)) \geq \operatorname{dist}(u, H^{-1}(\widehat{w})) \quad (u \in \mathcal{U}). \tag{4.8}$$

We refer to [11, 14, 17, 18, 21] for more on error bounds and metric subregularity. We need a partial version.

Definition 4.2. Let U, W be Hilbert spaces. Also let $M, P \in \mathcal{L}(U;U)$ and $N \in \mathcal{L}(W;W)$ with $N \geq 0$, $M \geq 0$, and $M \geq P$. We say say that $T : U \rightrightarrows W$ is *(P,N,M)-partially subregular* at $(\widehat{u}, \widehat{w}) \in \operatorname{Graph} T$ if there exists a neighbourhood $\mathcal{U} \ni \widehat{u}$ such that

$$\operatorname{dist}^2_N(\widehat{w}, T(u)) + \operatorname{dist}^2_{M-P}(u, T^{-1}(\widehat{w})) \geq \operatorname{dist}^2_M(u, T^{-1}(\widehat{w})) \quad (u \in \mathcal{U}). \tag{PSR}$$

We say that T is (N,M)-subregular if $P = M$.

Lemma 4.9. *Suppose $Z_{i+1}M_{i+1} \geq 0$ is self-adjoint and positive definite. Then*

$$\frac{1}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}} \geq \frac{1}{2}\operatorname{dist}^2_{Z_{i+1}(Z_{i+1}M_{i+1})^{-1}Z_{i+1}}(0, \widetilde{H}_{i+1}(u^{i+1})).$$

Proof. Let $q^{i+1} := -M_{i+1}(u^{i+1}-u^i)$. Then $q^{i+1} \in \widetilde{H}_{i+1}(u^{i+1})$. By applying $\frac{1}{2}\langle \cdot, u^{i+1}-u^i\rangle_{Z_{i+1}}$ to (PP, p.11), we therefore obtain

$$\frac{1}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}} = -\frac{1}{2}\langle q^{i+1}, u^{i+1}-u^i\rangle_{Z_{i+1}}.$$

By our assumptions $Z_{i+1}M_{i+1}$ is invertible. Therefore we can solve $u^{i+1}-u^i = -(Z_{i+1}M_{i+1})^{-1}Z_{i+1}q^{i+1}$. It follows

$$\frac{1}{2}\|u^{i+1}-u^i\|^2_{Z_{i+1}M_{i+1}} = \frac{1}{2}\|q^{i+1}\|^2_{Z^*_{i+1}(Z_{i+1}M_{i+1})^{-1}Z_{i+1}}.$$

This immediately yields the claim. □

As a consequence, for the linearly preconditioned case $V'_{i+1} = 0$ we obtain:

Proposition 4.10. *Suppose $Z_{i+1}M_{i+1}$ is self-adjoint and positive definite, and $V'_{i+1} = 0$. Let $N_{i+1} := \delta W^*_{i+1}Z^*_{i+1}(Z_{i+1}M_{i+1})^{-1}Z_{i+1}W_{i+1}$ for some $\delta \in [0,1]$. Let $P_{i+1} \in \mathcal{L}(U;U)$ with $Z_{i+2}M_{i+2} \geq Z_{i+1}P_{i+1}$. Then the partial error bound (PEB) holds if H is $(Z_{i+1}P_{i+1}, N_{i+1}, Z_{i+2}M_{i+2})$-partially subregular at some $\widehat{u} \in H^{-1}(0)$ in a neighbourhood $\mathcal{U}$ containing $\{u^i\}_{i=0}^{\infty}$.*

Exercise 4.2 (Basic proximal point method, subregularity). *Suppose for some* $\tau > 0, \pi \geq 0$, *and* $\delta \in [0,1]$ *that* H *is* $(\pi I, \delta\tau^2 I, (1+\pi)I)$*-partially subregular at* $(\widehat{u}, 0) \in \operatorname{Graph} H$, *and* $\langle H(u), u - \widehat{u}\rangle \geq 0$ *for all* $u \in U$. *The former is to say,*

$$\delta\tau^2 \operatorname{dist}^2(0, H(u)) + \operatorname{dist}^2(u, H^{-1}(0)) \geq (1+\pi)\operatorname{dist}^2(u, H^{-1}(0)) \quad (u \in \mathcal{U}).$$

Show that when this subregularity holds, the basic proximal point method $0 \in H(u^{i+1}) + \tau^{-1}(u^{i+1} - u^i)$ *satisfies*

$$\operatorname{dist}^2(u^N; H^{-1}(0)) \leq (1+\pi)^{-N} \operatorname{dist}^2(u^0; H^{-1}(0)).$$

For various examples, we refer to [30]. Although saddle point problems are also analysed there, it is still open whether we can derive practically useful results for them.

A NOTATION

We use $\Gamma(X)$ to denote the space of convex, proper, lower semicontinuous functions from X to the extended reals $\overline{\mathbb{R}} := [-\infty, \infty]$. We write ∂f for the convex subdifferential.

If $C \subset X$ is a convex set, we write

$$\delta_C(x) := \begin{cases} 0, & x \in C, \\ \infty, & x \notin C, \end{cases}$$

for the indicator function, and $N_C(x) = \partial \delta_C(x)$ for the normal cone at $x \in C$.

We use $\mathcal{L}(X; Y)$ to denote the space of bounded linear operators between Hilbert spaces X and Y. We denote the identity operator by I.

For $T, S \in \mathcal{L}(X; X)$, we write $T \geq S$ when $T - S$ is positive semidefinite.

Also for possibly non-self-adjoint $T \in \mathcal{L}(X; X)$, we introduce the inner product and norm-like notations

$$\langle x, z \rangle_T := \langle Tx, z \rangle, \quad \text{and} \quad \|x\|_T := \sqrt{\langle x, x \rangle_T}.$$

For $A \subset X$ a set, and $x \in X$, we write the distance to the set

$$\operatorname{dist}_T(x, A) := \inf_{x' \in A} \|x - x'\|_T.$$

For a set $A \subset \mathbb{R}$, we write $A \geq 0$ if every element $t \in A$ satisfies $t \geq 0$.

BIBLIOGRAPHY

[1] Aubin & Frankowska, *Set-Valued Analysis*, Springer, 1990.

[2] Bauschke & Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2017, DOI: 10.1007/978-3-319-48311-5.

[3] Benning, Knoll, Schönlieb & Valkonen, Preconditioned ADMM with nonlinear operator constraint, in: *System Modeling and Optimization: 27th IFIP TC 7 Conference, CSMO 2015, Sophia Antipolis, France, June 29–July 3, 2015, Revised Selected Papers*, Springer International Publishing, 2016, 117–126, DOI: 10.1007/978-3-319-55795-3_10, ARXIV: 1511.00425,

[4] Browder, Nonexpansive nonlinear operators in a Banach space, *Proceedings of the National Academy of Sciences of the United States of America* 54 (1965), 1041, URL: www.jstor.org/stable/73047.

[5] Chambolle & 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.

[6] Chambolle & Pock, On the ergodic convergence rates of a first-order primal–dual algorithm, *Mathematical Programming* (2015), 1–35, DOI: 10.1007/s10107-015-0957-3.

[7] Clarke, *Optimization and Nonsmooth Analysis*, Society for Industrial & Applied Mathematics, 1990, DOI: 10.1137/1.9781611971309.

[8] Clason, Mazurenko & Valkonen, Acceleration and global convergence of a first-order primal–dual method for nonconvex problems, 2018, ARXIV: 1802.03347, URL: tuomov.iki.fi/m/nlpdhgm_redo.pdf.

[9] Clason & Valkonen, Primal-dual extragradient methods for nonlinear nonsmooth PDE-constrained optimization, *SIAM Journal on Optimization* 27 (2017), 1313–1339, DOI: 10.1137/16M1080859, ARXIV: 1606.06219,

[10] Clason & Valkonen, Stability of saddle points via explicit coderivatives of pointwise subdifferentials, *Set-valued and Variational Analysis* 25 (2017), 69–112, DOI: 10.1007/s11228-016-0366-7, ARXIV: 1509.06582,

[11] Dontchev & Rockafellar, *Implicit Functions and Solution Mappings: A View from Variational Analysis*, Springer New York, 2014, DOI: 10.1007/978-1-4939-1037-3.

[12] Douglas & Rachford, On the Numerical Solution of Heat Conduction Problems in Two and Three Space Variables, *Transactions of the American Mathematical Society* 82 (1956), 421–439, DOI: 10.2307/1993056.

[13] Gabay, Applications of the Method of Multipliers to Variational Inequalities, in: *Studies in Mathematics and its Applications*, North-Holland, 1983, 299–331.

[14] Gfrerer, First order and second order characterizations of metric subregularity and calmness of constraint set mappings, *SIAM Journal on Optimization* 21 (2011), 1439–1474, DOI: 10.1137/100813415.

[15] He & Yuan, Convergence Analysis of Primal-Dual Algorithms for a Saddle-Point Problem: From Contraction Perspective, *SIAM Journal on Imaging Sciences* 5 (2012), 119–149, DOI: 10.1137/100814494.

[16] Hiriart-Urruty & Lemaréchal, *Convex analysis and minimization algorithms I-II*, Springer, 1993.

[17] Ioffe, *Variational Analysis of Regular Mappings: Theory and Applications*, Springer International Publishing, 2017, DOI: 10.1007/978-3-319-64277-2.

[18] Kruger, Error bounds and metric subregularity, *Optimization* 64 (2015), 49–79, DOI: 10.1080/02331934.2014.938074.

[19] Loris & Verhoeven, On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty, *Inverse Problems* 27 (2011), 125007, DOI: 10.1088/0266-5611/27/12/125007.

[20] Mordukhovich, *Variational Analysis and Generalized Differentiation I: Basic Theory*, Springer Verlag, 2006, DOI: 10.1007/3-540-31247-1.

[21] Ngai & Théra, Error Bounds in Metric Spaces and Application to the Perturbation Stability of Metric Regularity, *SIAM Journal on Optimization* 19 (2008), 1–20, DOI: 10.1137/060675721.

[22] Opial, Weak convergence of the sequence of successive approximations for nonexpansive mappings, *Bulletin of the American Mathematical Society* 73 (1967), 591–597, DOI: 10.1090/S0002-9904-1967-11761-0.

[23] Robinson, Some continuity properties of polyhedral multifunctions, in: *Mathematical Programming at Oberwolfach*, Springer Berlin Heidelberg, 1981, 206–214, DOI: 10.1007/BFb0120929.

[24] Rockafellar, *Convex Analysis*, Princeton University Press, 1972.

[25] Rockafellar & Wets, *Variational Analysis*, Springer, 1998, DOI: 10.1007/978-3-642-02431-3.

[26] Valkonen, A primal-dual hybrid gradient method for non-linear operators with applications to MRI, *Inverse Problems* 30 (2014), 055012, DOI: 10.1088/0266-5611/30/5/055012, ARXIV: 1309.5032,

[27] Valkonen, Set-valued analysis and optimisation, Lecture Notes, 2015, URL: tuomov.iki.fi/m/svao.pdf.

[28] Valkonen, Block-proximal methods with spatially adapted acceleration (2017), arXiv: 1609.07373, URL: tuomov.iki.fi/m/blockcp.pdf.

[29] Valkonen, Optimisation for computer vision and data science, Lecture Notes, 2017, URL: tuomov.iki.fi/m/optvis.pdf.

[30] Valkonen, Preconditioned proximal point methods and notions of partial subregularity, 2017, arXiv: 1711.05123, URL: tuomov.iki.fi/m/subreg.pdf.

[31] Valkonen, Testing and non-linear preconditioning of the proximal point method, 2017, arXiv: 1703.05705, URL: tuomov.iki.fi/m/proxtest.pdf.

[32] Valkonen & Pock, Acceleration of the PDHGM on partially strongly convex functions, *Journal of Mathematical Imaging and Vision* 59 (2017), 394–414, DOI: 10.1007/s10851-016-0692-2, arXiv: 1511.06566,