Appendix D

Bibliographic Notes

Chapter 1. For a comprehensive treatment of finite-dimensional vector spaces and advanced linear algebra topics, the reader can refer to the classical book of Halmos [64], as well as to the textbooks of Meyer [86] and Strang [117].

Chapters 2, 3, 4. Most of the material in these chapters is classical. Additional materials and extensions can be found, for example, in Bauschke and Combettes [8], Bertsekas [29], Borwein and Lewis [32], Hiriart-Urruty and Lemaréchal [67], Nesterov [94] and Rockafellar [108]. Example 2.17 is taken from the book of Hiriart-Urruty and Lemaréchal [67, Example 2.1.4]. Example 2.32 is from Rockafellar [108, p. 83]. The proof in Example 3.31 follows Beck and Teboulle [20, Theorem 4.1]. Section 3.5, excluding Theorem 3.60, follows Hiriart-Urruty and Lemaréchal [67, Section VII.3.3]. Theorem 3.60 is a slight extension of Lemma 6 from Lan [78]. The optimality conditions derived in Example 3.66 are rather old and can be traced back to Sturm, who proved them in his work from 1884 [118]. Actually, (re)proving these conditions was the main motivation for Weiszfeld to devise the (now-called) Weiszfeld's method in 1937 [124]. For more information on the Fermat–Weber problem and Weiszfeld's method, see the review paper of Beck and Sabach [14] and references therein.

Chapter 5. The proof of the descent lemma can be found in Bertsekas [28]. The proof of Theorem 5.8 follows the proof of Nesterov in [94, Theorem 2.1.5]. The equivalence between claims (i) and (iv) in Theorem 5.8 is also known as the Baillon-Haddad theorem [5]. The analysis in Example 5.11 of the smoothness parameter of the squared l_p-norm follows the derivation in the work of Ben-Tal, Margalit, and Nemirovski [24, Appendix 1]. The conjugate correspondence theorem can be deduced from the work of Zalinescu [128, Theorem 2.2] and can also be found in the paper of Azé and Penot [3] as well as Zalinescu's book [129, Corollary 3.5.11]. In its Euclidean form, the result can be found in the book of Rockafellar and Wets [111, Proposition 12.60]. Further characterizations appear in the paper of Bauschke and Combettes [7]. The proof of Theorem 5.30 follows Beck and Teboulle [20, Theorem 4.1].

Chapter 6. The seminal 1965 paper of Moreau [87] already contains much of the properties of the proximal mapping discussed in the chapter. Excellent references for the subject are the book of Bauschke and Combettes [8], the paper of Combettes and Wajs [44], and the review paper of Parikh and Boyd [102]. The computation of the prox of the squared l_1-norm in Section 6.8.2 is due to Evgeniou, Pontil, Spinellis, and Nassuphis [54].

Chapter 7. The notion of symmetry w.r.t. a given set of orthogonal matrices was studied by Rockafellar [108, Chapter 12]. A variant of the symmetric conjugate theorem (Theorem 7.9) can be found in Rockafellar [108, Corollary 12.3.1]. Fan's inequality can be found in Theobald [119]. Von Neumann's trace inequality [123], as well as Fan's inequality, are often formulated over the complex field, but the adaptation to the real field is straightforward. Sections 7.2 and 7.3, excluding the spectral proximal theorem, are based on the seminal papers of Lewis [80, 81] on unitarily invariant functions. See also Borwein and Lewis [32, Section 1.2], as well as Borwein and Vanderwerff [33, Section 3.2]. The equivalence between the convexity of spectral functions and their associated functions was first established by Davis in [47]. The spectral proximal formulas can be found in Parikh and Boyd [102].

Chapter 8. Example 8.3 is taken from Vandenberghe's lecture notes [122]. Wolfe's example with $\gamma = \frac{16}{9}$ originates from his work [125]. The version with general $\gamma > 1$, along with the support form of the function, can be found in the set of exercises [35]. Studies of subgradient methods and extensions can be found in many books; to name a few, the books of Nemirovsky and Yudin [92], Shor [116] and Polyak [104] are classical; modern accounts of the subject can be found, for example, in Bertsekas [28, 29, 30], Nesterov [94], and Ruszczyński [113]. The analysis of the stochastic and deterministic projected subgradient method in the strongly convex case is based on the work of Lacoste-Julien, Schmidt, and Bach [77]. The fundamental inequality for the incremental projected subgradient is taken from Nedić and Bertsekas [89], where many additional results on incremental methods are derived. Theorem 8.42 and Lemma 8.47 are Lemmas 1 and 3 from the work of from Nedić and Ozdaglar [90]. The latter work also contains additional results on the dual projected subgradient method with constant stepsize. The presentation of the network utility maximization problem, as well as the distributed subgradient method for solving it, originates from Nedić and Ozdaglar [91].

Chapter 9. The mirror descent method was introduced by Nemirovsky and Yudin in [92]. The interpretation of the method as a non-Euclidean projected subgradient method was presented by Beck and Teboulle in [15]. The rate of convergence analysis of the mirror descent method is based on [15]. The three-points lemma was proven by Chen and Teboulle in [43]. The analysis of the mirror-C method is based on the work of Duchi, Shalev-Shwartz, Singer, and Tewari [49], where the algorithm is introduced in an online and stochastic setting.

Chapter 10. The proximal gradient method can be traced back to the forward-backward algorithm introduced by Bruck [36], Pasty [103], and Lions and Mercier [83]. More modern accounts of the topic can be found, for example, in Bauschke and Combettes [8, Chapter 27], Combettes and Wajs [44], and Facchinei and Pang

[55, Chapter 12]. The proximal gradient method is a generalization of the gradient method, which goes back to Cauchy [38] and was extensively studied and generalized by many authors; see, for example, the books of Bertsekas [28], Nesterov [94], Polyak [104], and Nocedal and Wright [99], as well as the many references therein. ISTA and its variations was studied in the literature in several contexts; see, for example, the works of Daubechies, Defrise, and De Mol [46]; Hale, Yin, and Zhang [63]; Wright, Nowak, and Figueiredo [127]; and Elad [52]. The analysis of the proximal gradient method in Sections 10.3 and 10.4 mostly follows the presentation of Beck and Teboulle in [18] and [19]. Lemma 10.11 was stated and proved for the case where g is an indicator of a nonempty closed and convex set in [9]; see also [13, Lemma 2.3]. Theorem 10.9 on the monotonicity of the gradient mapping is a simple generalization of [10, Lemma 9.12]. The first part of the monotonicity result was shown in the case where g is an indicator of a nonempty closed and convex set in Bertsekas [28, Lemma 2.3.1]. Lemma 10.12 is a minor variation of Lemma 2.4 from Necoara and Patrascu [88]. Theorem 10.26 is an extension of a result of Nesterov from [97] on the convergence of the gradient method for convex functions. The proximal point method was studied by Rockafellar in [110], as well as by many other authors; see, for example, the book of Bauschke and Combettes [8] and its extensive list of references. FISTA was developed by Beck and Teboulle in [18]; see also the book chapter [19]; the convergence analysis presented in Section 10.7 is taken from these sources. When the nonsmooth part is an indicator function of a closed and convex set, the method reduces to the optimal gradient method of Nesterov from 1983 [93]. Other accelerated proximal gradient methods can be found in the works of Nesterov [98] and Tseng [121]—the latter also describes a generalization to the non-Euclidean setting, which is an extension of the work of Auslender and Teboulle [2]. MFISTA and its convergence analysis are from the work of Beck and Teboulle [17]. The idea of using restarting in order to gain an improved rate of convergence in the strongly convex case can be found in Nesterov's work [98] in the context of a different accelerated proximal gradient method, but the idea works for any method that gains an $O(1/k^2)$ rate in the (not necessarily strongly) convex case. The proof of Theorem 10.42 follows the proof of Theorem 4.10 from the review paper of Chambolle and Pock [42]. The idea of solving nonsmooth problems through a smooth approximation was studied by many authors; see, for example, the works of Ben-Tal and Teboulle [25], Bertsekas [26], Moreau [87], and the more recent book of Auslender and Teboulle [1] and references therein. Lemma 10.70 can be found in Levitin and Polyak [79]. The idea of producing an $O(1/\varepsilon)$ complexity result for nonsmooth problems by employing an accelerated gradient method was first presented and developed by Nesterov in [95]. The extension to the three-part composite model and to the setting of more general smooth approximations was studied by Beck and Teboulle [20], where additional results and extensions can also be found. The non-Euclidean gradient method was proposed by Nutini, Schmidt, Laradji, Friendlander, and Koepke [100], where its rate of convergence in the strongly convex case was analyzed; the work [100] also contains a comparison between two coordinate selection strategies: Gauss–Southwell (which is the one considered in the chapter) and randomized selection. The non-Euclidean proximal gradient method was presented in the work of Tseng [121], where an accelerated non-Euclidean version was also analyzed.

Chapter 11. The version of the block proximal gradient method in which the nonsmooth functions g_i are indicators was studied by Luo and Tseng in [84], where some error bounds on the model were assumed. It was shown that under the model assumptions, the CBPG method with each block consisting of a single variable has a linear rate of convergence. Nesterov studied in [96] a randomized version of the method (again, in the setting where the nonsmooth functions are indicators) in which the selection of the block on which a gradient projection step is performed at each iteration is done randomly via a pre-described distribution. For the first time, Nesterov was able to establish global nonasymptotic rates of convergence in the convex case without any strict convexity, strong convexity, uniqueness, or error bound assumptions. Specifically, it was shown that the rate of convergence to the optimal value of the expectation sequence of the function values of the sequence generated by the randomized method is sublinear under the assumption of Lipschitz continuity of the gradient and linear under a strong convexity assumption. In addition, an accelerated $O(1/k^2)$ was devised in the unconstrained setting. Probabilistic results on the convergence of the function values were also provided. In [107] Richtarik and Takac generalized Nesterovs results to the composite model. The derivation of the randomized complexity result in Section 11.5 mostly follows the presentation in the work of Lin, Lu, and Xiao [82]. The type of analysis in the deterministic convex case (Section 11.4.2) originates from Beck and Tetruashvili [22], who studied the case in which the nonsmooth functions are indicators. The extension to the general composite model can be found in Shefi and Teboulle [115] as well as in Hong, Wang, Razaviyayn, and Luo [69]. Lemma 11.17 is Lemma 3.8 from [11]. Theorem 11.20 is a specialization of Lemma 2 from Nesterov [96]. Additional related methods and discussions can be found in the extensive survey of Wright [126].

Chapter 12. The idea of using a proximal gradient method on the dual of the main model (12.1) was originally developed by Tseng in [120], where the algorithm was named "alternating minimization." The primal representations of the DPG and FDPG methods, convergence analysis, as well as the primal-dual relation are from Beck and Teboulle [21]. The DPG method for solving the total variation problem was initially devised by Chambolle in [39], and the accelerated version was considered by Beck and Teboulle [17]. The one-dimensional total variation denoising problem is presented as an illustration for the DPG and FDPG methods; however, more direct and efficient methods exist for tackling the problem; see Hochbaum [68], Condat [45], Johnson [73], and Barbero and Sra [6]. The dual block proximal gradient method was discussed in Beck, Tetruashvili, Vaisbourd, and Shemtov [23], from which the specific decomposition of the isotropic two-dimensional total variation function is taken. The accelerated method ADBPG is a different representation of the accelerated method proposed by Chambolle and Pock in [41]. The latter work also discusses dual block proximal gradient methods and contains many other suggestions for decompositions of total variation functions.

Chapter 13. The conditional gradient algorithm was presented by Frank and Wolfe [56] in 1956 for minimizing a convex quadratic function over a compact polyhedral set. The original paper of Frank and Wolfe also contained a proof of an $O(1/k)$ rate of convergence in function values. Levitin and Polyak [79] showed that this $O(1/k)$ rate can also be extended to the case where the feasible set is a general compact con-

vex set and the objective function is L-smooth and convex. Dunn and Harshbarger [50] were probably the first to suggest a diminishing stepsize rule for the conditional gradient method and to establish a sublinear rate under such a strategy. The generalized conditional gradient method was introduced and analyzed by Bach in [4], where it was shown that under a certain setting, it can be viewed as a dual mirror descent method. Lemma 13.7 (fundamental inequality for generalized conditional gradient) can be found in the setting of the conditional gradient method in Levitin and Polyak [79]. The interpretation of the power method as the conditional gradient method was described in the work of Luss and Teboulle [85], where many other connections between the conditional gradient method and the sparse PCA problem are explored. Lemma 13.13 is an extension of Lemma 4.4 from Bach's work [4], and the proof is almost identical. Similar results on sequences of nonnegative numbers can be found in the book of Polyak [104, p. 45]. Section 13.3.1 originates from the work of Canon and Cullum [37]. Polyak in [104, p. 214, Exercise 10] seems to be the first to mention the linear rate of convergence of the conditional gradient method under a strong convexity assumption on the feasible set. Theorem 13.23 is from Journée, Nesterov, Richtárik, and Sepulchre [74, Theorem 12]. Lemma 13.26 and Theorem 13.27 are from Levitin and Polyak [79], and the exact form of the proof is due to Edouard Pauwels. Another situation, which was not discussed in the chapter, in which linear rate of converge can be established, is when the objective function is strongly convex and the optimal solution resides in the interior of the feasible set (Guélat and Marcotte [62]). Epelman and Freund [53], as well as Beck and Teboulle [16], showed a linear rate of convergence of the conditional gradient method with a special stepsize choice in the context of finding a point in the intersection of an affine space and a closed and convex set under a Slater-type assumption. The randomized generalized block conditional gradient method presented in Section 13.4 is a simple generalization of the randomized block conditional gradient method introduced and analyzed by Lacoste-Julien, Jaggi, Schmidt, and Pletscher in [76]. A deterministic version was analyzed by Beck, Pauwels, and Sabach in [12]. An excellent overview of the conditional gradient method, including many more theoretical results and applications, can be found in the thesis of Jaggi [72].

Chapter 14. The alternating minimization method is a rather old and fundamental algorithm. It appears in the literature under various names such as the block-nonlinear Gauss-Seidel method or the block coordinate descent method. Powell's example appears in [106]. Theorem 14.3 and its proof originate from Bertsekas [28, Proposition 2.7.1]. Theorem 14.9 and its proof are an extension of Proposition 6 from Grippo and Sciandrone [61] to the composite model. The proof of Theorem 14.11 follows the proof of Theorem 3.1 from the work of Hong, Wang, Razaviyayn, and Luo [69], where more general schemes than alternating minimization are also considered. Section 14.5.2 follows [11].

Chapter 15 The augmented Lagrangian method can be traced back to Hestenes [66] and Powell [105]. The method and its many variants was studied extensively in the literature, see, for example, the books of Bertsekas [27] and Bertsekas and Tsitsiklis [31] and references therein. Rockafellar [109] was first to establish the duality between the proximal point and the augmented Lagrangian methods; see also additional discussions in the work of Iusem [71]. ADMM is equivalent to an

operator splitting method called Douglas–Rachford splitting, which was introduced in the 1950s for the numerical solution of partial differential equations [48]. ADMM, as presented in the chapter, was first introduced by Gabay and Mercier [57] and Glowinski and Marrocco [59]. An extremely extensive survey on ADMM method can be found in the work of Boyd, Parikh, Chu, Peleato, and Eckstein [34]. AD-PMM was suggested by Eckstein [51]. The proof of Theorem 15.4 on the rate of convergence of AD-PMM is based on a combination of the proof techniques of He and Yuan [65] and Gao and Zhang [58]. Shefi and Teboulle provided in [114] a unified analysis for general classes of algorithm that include AD-PMM as a special instance. Shefi and Teboulle also showed the relation between AD-LPMM and the Chambolle–Pock algorithm [40].