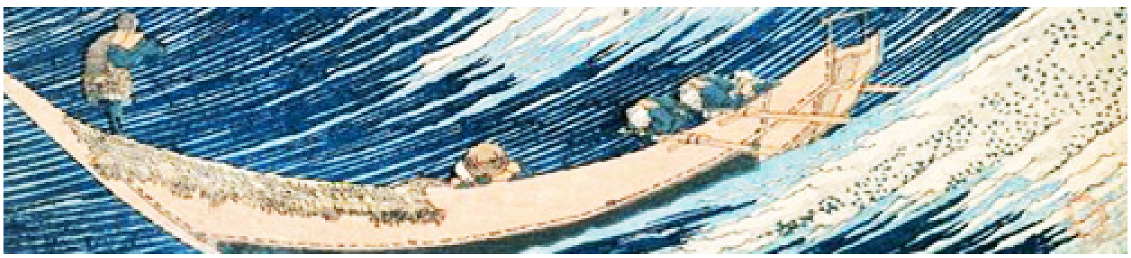
Mathematical Foundations of Data Sciences



Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
www.gpeyre.com
www.numerical-tours.com

October 19, 2017

Chapter 12

Convex Analysis

The main references for this chapter are [14, 5]. This chapters uses different notations than the previous one, and we denote $f(x)$ a typical function to minimized with respect to the variable x. We discus here some important concepts from convex analysis and use them to study theoretically the performance of ℓ^1-type methods.

12.1 Basics of Convex Analysis

We consider minimization problems of the form

$$\min_{x\in\mathcal{H}} f(x) \tag{12.1}$$

over the finite dimension (Hilbertian) space $\mathcal{H} \overset{\text{def.}}{=} \mathbb{R}^N$, with the canonical inner product $\langle \cdot, \cdot \rangle$. Most of the results of this chapter extends to possibly infinite dimensional Hilbert space.

Here $f : \mathcal{H} \longrightarrow \bar{\mathbb{R}} \overset{\text{def.}}{=} \mathbb{R} \cup \{+\infty\}$ is a convex function. Note that we allow here f to take the value $+\infty$ to integrate constraints in the objective, and the constraint set is thus the "domain" of the function

$$\operatorname{dom}(f) \overset{\text{def.}}{=} \{x \;;\; f(x) < +\infty\}.$$

A useful notation is the indicator function of a set $\mathcal{C} \subset \mathcal{H}$

$$\iota_{\mathcal{C}}(x) \overset{\text{def.}}{=} \begin{cases} 0 & \text{if} \quad x \in \mathcal{C}, \\ +\infty & \text{otherwise.} \end{cases}$$

12.1.1 Convex Sets and Functions

A convex set $\Omega \subset \mathcal{H}$ is such that

$$\forall\, (x, y, t) \in \mathcal{H}^2 \times [0, 1], \quad (1-t)x + ty \in \Omega.$$

A convex function is such that

$$\forall\, (x, y, t) \in \mathcal{H}^2 \times [0, 1], \quad f((1-t)x + ty) \leqslant (1-t)f(x) + tf(y) \tag{12.2}$$

and this is equivalent to its epigraph $\{(x, r) \in \mathcal{H} \times \mathbb{R} \;;\; r \geqslant f(x)\}$ being a convex set. Note that here we use $\leqslant$ as a comparison over $\bar{\mathbb{R}}$. The function f is strictly convex if equality in (12.2) only holds for $t \in \{0, 1\}$. A set Ω being convex is equivalent to $\iota_{\mathcal{C}}$ being a convex function.

In the remaining part of this chapter, we consider convex function f which are proper, i.e. such that $\operatorname{dom}(f) \neq \emptyset$, and that should be lower-semi-continuous (lsc), i.e. such that for all $x \in \mathcal{H}$,

$$\liminf_{y \to x} f(y) \geqslant f(x).$$

It is equivalent to $\operatorname{epi}(f)$ being a closed convex set. We denote $\Gamma_0(\mathcal{H})$ the set of proper convex lsc functions.

12.1.2 First Order Conditions

Existence of minimizers. Before looking at first optimality conditions, one has to check that there exists minimizers, which is implied by the l.s.c. property and coercivity.

Proposition 25. *If f is l.s.c. and coercive (i.e. $f(x) \to +\infty$ as $x \to +\infty$), then there exists a minimizer $x^\star$ of f.*

Proof. Since f is coercive, it is bounded from bellow, one can consider a minimizing sequence $(x_n)_n$ such that $f(x_n) \to \min f$. Since f is l.s.c., this implies that the sub-level set of f are closed, and coercivity imply they are bounded, hence compact. One can thus extract from $(x_n)_n$ a converging sub-sequence $(x_{n(p)})_p$, $x_{n(p)} \to x^\star$. Lower semi-continuity implies that $\min f = \lim_p f(x_{n(p)}) \geqslant f(x^\star)$, and hence $x^\star$ is a minimizer. $\square$

This existence proof is often called the "direct method of calculus of variation". Note that if the function f is in $\Gamma_0(\mathcal{H})$, then the set of minimizer $\operatorname{argmin} f$ is a closed convex set, and all local minimizers (i.e. minimizer of the function restricted to an open ball) are global one. If it is furthermore strictly convex, then there is a single minimizer.

Sub-differential. The sub-differential at x of such a f is defined as

$$\partial f(x) \stackrel{\text{def.}}{=} \{u \in \mathcal{H}^* \; ; \; \forall y, f(y) \geqslant f(x) + \langle u, \, z - x \rangle\} \,.$$

We denote here $\mathcal{H}^* = \mathbb{R}^N$ the set of "dual" vector. Although in finite dimensional Euclidean space, this distinction is not needed, it helps to distinguish primal from dual vectors, and recall that the duality pairing implicitly used depends on the choice of an inner product. The sub-differential $\partial f(x)$ is thus the set of "slopes" u of tangent affine planes $f(x) + \langle u, \, z - x \rangle$ that fits bellow the graph of f.

Note that f being differentiable at x is equivalent to the sub-differential being reduced to a singleton (equal to the gradient vector)

$$\partial f(x) = \{\nabla f(x)\}.$$

Informally, the "size" of $\partial f(x)$ controls how smooth f is at x.

One easily checks that $\partial f(x) \subset \mathcal{H}^*$ is a convex set, and it is non-empty if and only if $x \in \operatorname{dom}(f)$. The operator $\partial f : \mathcal{H} \mapsto 2^{\mathcal{H}^*}$ is thus "set-valued", and we often denote this as $\partial f : \mathcal{H} \hookrightarrow \mathcal{H}^*$.

Remark 6 (Maximally monotone operator). The operator ∂f is particular instance of so-called monotone operator, since one can check that $U = \partial f$ satisfies

$$\forall \, (u, v) \in U(x) \times U(y), \quad \langle y - x, \, v - u \rangle \geqslant 0.$$

In the 1-D setting, being monotone is the same as being an increasing map. Sub-differential can also be shown to be maximally monotone, in the sense that such an operator is not striclty included in the graph of another monotone operator. Note that there exists monotone maps which are not subdifferential, for instance $(x, y) \mapsto (-y, x)$. Much of the theory of convex analysis and convex optimization can be extended to deal with arbitrary maximally monotone-maps in place of subdifferential, but we will not pursue this here.

A prototypical example is the absolute value $f(x) = |\cdot|$, and writing conveniently $\partial f(x) = \partial|\cdot|(x)$, one verifies that

$$\partial|\cdot|(x) = \begin{cases} -1 & \text{if} \quad x < 0, \\ +1 & \text{if} \quad x > 0, \\ [-1, 1] & \text{if} \quad x = 0. \end{cases}$$

[ToDo: sub-differential of an indicator]

Sub-differential calculus. There is a large set of calculus rules that allows to simplify the computation of sub-differentials. For decomposable function $f(x_1,\ldots,x_K) = \sum_{k=1}^K f_k(x_k)$, the sub-differential is the product of the sub-differentials

$$\partial f(x_1,\ldots,x_K) = \partial f_1(x_1) \times \ldots \times \partial f_K(x_K).$$

This can be used to compute the sub-differential of the ℓ^1 norm $\|x\|_1 = \sum_{k=1}^N |x_k|$

$$\partial \|\cdot\|_1(x) = \prod_{k=1}^N \partial|\cdot|(x_k)$$

which is thus an hyper rectangle. This means that, denoting $I = \text{supp}(x)$, one has $u \in \partial\|\cdot\|_1(x)$ is equivalent to

$$u_I = \text{sign}(x_I) \quad \text{and} \quad \|u_{I^c}\|_\infty \leqslant 1.$$

A tricky problem is to compute the sub-differential of the sum of two functions. If one of the two function is continuous at x (i.e. it has a finite value), then

$$\partial(f+g)(x) = \partial f(x) \oplus \partial g(x) = \{u+v \; ; \; (u,v) \in \partial f(x) \times \partial g(x)\}$$

where $\oplus$ thus denotes the Minkowski sum. For instance, if f is differentiable at x, then

$$\partial(f+g)(x) = \nabla f(x) + \partial g(x) = \{\nabla f(x) + v \; ; \; v \in \partial g(x)\}.$$

Positive linear scaling is simple to handle

$$\forall \lambda \in \mathbb{R}_+, \quad \partial(\lambda f)(x) = \lambda(\partial f(x)).$$

The chain rule for sub-differential is difficult since in general composition does not work so-well with convexity. The only simple case is composition with linear functions, which preserves convexity. Denoting $A \in \mathbb{R}^{P\times N}$ and $f \in \Gamma_0(\mathbb{R}^P)$, one has that $f \circ A \in \Gamma_0(\mathbb{R}^N)$ and

$$\partial(f \circ A)(x) = A^*(\partial f)(Ax) \overset{\text{def.}}{=} \{A^* u \; ; \; u \in \partial f(Ax)\}.$$

12.2 Convex Duality

Duality is associated to a particular formulation of the optimization problem, so that for instance making change of variables results in a different duality.

12.2.1 Fenchel Transform

12.2.2 Lagrange Duality

12.2.3 Fenchel-Rockafeller Duality

Bibliography

[1] P. Alliez and C. Gotsman. Recent advances in compression of 3d meshes. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, editors, *Advances in multiresolution for geometric modelling*, pages 3–26. Springer Verlag, 2005.

[2] P. Alliez, G. Ucelli, C. Gotsman, and M. Attene. Recent advances in remeshing of surfaces. In *AIM@SHAPE repport*. 2005.

[3] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, 2014.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[5] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[6] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun. on Pure and Appl. Math.*, 57(2):219–266, 2004.

[7] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Académie des Sciences*, Serie I(346):589–592, 2006.

[8] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Modeling and Simulation*, 5:861–899, 2005.

[9] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20:89–97, 2004.

[10] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.

[11] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[12] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[13] F. R. K. Chung. Spectral graph theory. *Regional Conference Series in Mathematics, American Mathematical Society*, 92:1–212, 1997.

[14] Philippe G Ciarlet. Introduction à l'analyse numérique matricielle et à l'optimisation. 1982.

[15] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simulation*, 4(4), 2005.

[16] P. Schroeder et al. D. Zorin. Subdivision surfaces in character animation. In *Course notes at SIGGRAPH 2000*, July 2000.

[17] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. on Pure and Appl. Math.*, 57:1413–1541, 2004.

[18] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.*, 4(3):245–267, 1998.

[19] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, Dec 1994.

[20] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

[21] M. Figueiredo and R. Nowak. An EM Algorithm for Wavelet-Based Image Restoration. *IEEE Trans. Image Proc.*, 12(8):906–916, 2003.

[22] M. S. Floater and K. Hormann. Surface parameterization: a tutorial and survey. In N. A. Dodgson, M. S. Floater, and M. A. Sabin, editors, *Advances in multiresolution for geometric modelling*, pages 157–186. Springer Verlag, 2005.

[23] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

[24] I. Guskov, W. Sweldens, and P. Schröder. Multiresolution signal processing for meshes. In Alyn Rockwood, editor, *Proceedings of the Conference on Computer Graphics (Siggraph99)*, pages 325–334. ACM Press, August8–13 1999.

[25] A. Khodakovsky, P. Schröder, and W. Sweldens. Progressive geometry compression. In *Proceedings of the Computer Graphics Conference 2000 (SIGGRAPH-00)*, pages 271–278, New York, July 23–28 2000. ACMPress.

[26] L. Kobbelt. $\sqrt{3}$ subdivision. In Sheila Hoffmeyer, editor, *Proc. of SIGGRAPH'00*, pages 103–112, New York, July 23–28 2000. ACMPress.

[27] M. Lounsbery, T. D. DeRose, and J. Warren. Multiresolution analysis for surfaces of arbitrary topological type. *ACM Trans. Graph.*, 16(1):34–73, 1997.

[28] S. Mallat. *A Wavelet Tour of Signal Processing, 3rd edition.* Academic Press, San Diego, 2009.

[29] Stephane Mallat. *A wavelet tour of signal processing: the sparse way.* Academic press, 2008.

[30] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. on Pure and Appl. Math.*, 42:577–685, 1989.

[31] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[32] Gabriel Peyré. *L'algèbre discrète de la transformée de Fourier.* Ellipses, 2004.

[33] Gabriel Peyré and Marco Cuturi. Computational optimal transport. 2017.

[34] J. Portilla, V. Strela, M.J. Wainwright, and Simoncelli E.P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Proc.*, 12(11):1338–1351, November 2003.

[35] E. Praun and H. Hoppe. Spherical parametrization and remeshing. *ACM Transactions on Graphics*, 22(3):340–349, July 2003.

[36] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[37] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 2015.

[38] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, Frank Lenzen, and L Sirovich. *Variational methods in imaging.* Springer, 2009.

[39] P. Schröder and W. Sweldens. Spherical Wavelets: Efficiently Representing Functions on the Sphere. In *Proc. of SIGGRAPH 95*, pages 161–172, 1995.

[40] P. Schröder and W. Sweldens. Spherical wavelets: Texture processing. In P. Hanrahan and W. Purgathofer, editors, *Rendering Techniques '95.* Springer Verlag, Wien, New York, August 1995.

[41] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[42] A. Sheffer, E. Praun, and K. Rose. Mesh parameterization methods and their applications. *Found. Trends. Comput. Graph. Vis.*, 2(2):105–171, 2006.

[43] Jean-Luc Starck, Fionn Murtagh, and Jalal Fadili. *Sparse image and signal processing: Wavelets and related geometric multiscale analysis.* Cambridge university press, 2015.

[44] W. Sweldens. The lifting scheme: A custom-design construction of biorthogonal wavelets. *Applied and Computation Harmonic Analysis*, 3(2):186–200, 1996.

[45] W. Sweldens. The lifting scheme: A construction of second generation wavelets. *SIAM J. Math. Anal.*, 29(2):511–546, 1997.