Do Androids Dream of Henri Poincaré with Hierarchical Optimization ?

Optimization using subliminal mechanism ?

Mathematical optimization has been a major driving force of modern AI technologies and data sciences (see, e.g., [1,2,3]). A question arises: **what is a groundbreaking optimization model that is expected to bring about dramatical evolution in next-generation AI ?** A valuable hint for this visionary question could be found only if we try to reveal the ingenious ways of thinking of exceptionally gifted humans, e.g., great mathematicians and grand masters of board games. In my student days long ago, I found, in *Mathematical Discovery* by Henri Poincaré (see, e.g., Chapter III of [4, pp.387-400]), the following impressive words: (i) *Everything happens as if the discoverer were a secondary examiner who had only to interrogate candidates declared eligible after passing a preliminary test* [4, p.391], (ii) *Of the very large number of combinations which the subliminal ego blindly forms, almost all are without interest and without utility. But, for that very reason, they are without action on the aesthetic sensibility; the conscious will never know them* [4, p.397]. Poincaré's words seem for me now to suggest that breakthrough ideas represented by outstanding mathematical discovery can be achieved through some mysterious process of double stage search where the first stage search is performed with the aid of a certain aesthetic sensibility in unconscious field. Similar words are also found in Yoshiharu Habu's explanation [5] on his aesthetic sensibility that unconsciously helps him select a breakthrough move in a crucial phase of shogi game.

Their explanations tempt me to formulate a hypothesis that their brains are exploiting simultaneously two different criteria, say Φ for the preliminary tests in their unconscious fields and Ψ for the secondary tests in their conscious ones and to model their ingenious search as a certain computational process for the **hierarchical optimization**:

$$\left.\begin{array}{c}\text{minimize } \Psi(\mathbf{x}^\star) \\ \text{subject to } \mathbf{x}^\star \in \mathcal{S}_\star := \underset{\mathbf{x}\in\mathbb{R}^N}{\operatorname{argmin}}\, \Phi(\mathbf{x})\end{array}\right\} \quad (1)$$

rather than the traditional optimization model just for minimization of Φ, where the function Ψ is newly introduced for the second stage optimization. My amateur hypothesis does not contradict Sigmund Freud's psychoanalysis [6] saying that information stored in unconscious field of our brain has various level of difficulty for transforming it into available form in conscious field. Such a transform seems to correspond to the selection process explained by Poincaré and Habu in terms of aesthetic sensibility (Note: A similar consideration on the role of sub-conscious representation for creativity is found [7] but not in the context of optimization models). My naive imagination toward **optimization using subliminal mechanism** has also been motivated by studies on the neural associative memory [8] and the resting state fMRI [9].

Professor Isao Yamada

PhD, FIEEE

Technical co-chairs of APSIPA-ASC 2018

Professor

Department of Information and Communications Engineering,

Tokyo institute of Technology

Isao Yamada is a professor with the Department of Information and Communications Engineering, and the Director of the Global Scientific Information and Computing Center, Tokyo Institute of Technology. His current research interests are in mathematical signal processing, machine learning, nonlinear inverse problems, and optimization theory. He has been a Fellow of IEEE and IEICE since 2015. He received the MEXT Minister Award (Research Category) in 2016, the IEEE Signal Processing Magazine Best Paper Award in 2015, the IEICE Achievement Award in 2009, and the Docomo Mobile Science Award (Fundamental Science Division) in 2005.

What can we do for hierarchical optimization ?

The hierarchical optimization in (1) seems to be our ideal target but in reality the computation of its solution must be very challenging, as suggested by Tikhonov approximation theorem [10], even if Φ and Ψ are convex functions. To keep the currently achievable applicability by the state-of-the-art non-hierarchical convex optimization algorithms, we model the first stage cost function in (1) as

$$\Phi(\mathbf{x}) := f(\mathbf{x}) + \sum_{i=1}^{m} g_i(A_i\mathbf{x}), \quad (2)$$

where $f\colon \mathbb{R}^N \to (-\infty,\infty]$ and $g_i\colon \mathbb{R}^{N_i} \to (-\infty,\infty]$ $(i = 1,2,\dots,m)$ are convex but not necessarily differentiable everywhere, and $A_i \in \mathbb{R}^{N_i\times N}$ $(i = 1,2,\dots,m)$. Fortunately, a unified perspective from the view point of convex analysis and monotone operator theory (see, e.g., [11,12]) tells us that many convex optimization scenarios in data sciences, machine learning, and signal processing, appear as instances of the model (2) and that the so-called **proximity operators** of f and g_i $(i = 1,2,\dots,m)$ are available [11] as building blocks [13] of a computable

nonexpansive operator $T: \mathcal{H} \to \mathcal{H}$ and a bounded linear operator $\Xi: \mathcal{H} \to \mathbb{R}^N$ satisfying

$$\mathcal{S}_\star = \underset{\mathbf{x}\in\mathbb{R}^N}{\operatorname{argmin}}\ \Phi(\mathbf{x}) = \{\Xi(\mathbf{z}) \in \mathbb{R}^N \mid \mathbf{z} \in \operatorname{Fix}(T)\},$$

where $\mathcal{H}$ is a certain real Hilbert space, not necessarily $\mathcal{H} = \mathbb{R}^N$, and $\operatorname{Fix}(T) \coloneqq \{\boldsymbol{z} \in \mathcal{H} \mid T(\boldsymbol{z}) = \boldsymbol{z}\}$ is the set of all fixed points of T. Indeed, by plugging the nonexpansive operator T and the convex function $\Theta \coloneqq \Psi \circ \Xi$ into the **hybrid steepest descent method** [13,14]:

$$\mathbf{z}_{n+1} = T(\mathbf{z}_n) - \lambda_{n+1}\nabla\Theta(T(\mathbf{z}_n)) \qquad (3)$$

with a slowly decreasing sequence $(\lambda_n)_{n\geq 1} \subset [0, \infty)$, under reasonable conditions, we can generate a sequence $\Xi(\mathbf{z}_n)$ $(n = 0,1,2,\dots)$ which converges to a solution of (1).

An application to Cortes-Vapnik problem

To demonstrate the inherent applicability of the hierarchical convex optimization to machine learning problems, I conclude this article with a short introduction on our recent application [13] to a novel hierarchical convex relaxation of the Cortes-Vapnik problem [15, Sec.3]. This application has been made for sound extension of a central idea in the classical Support Vector Machine (SVM) [1] to be applicable to general training dataset:

$$\mathcal{D} \coloneqq \{(\mathbf{x}_i, \mathfrak{L}(\mathbf{x}_i)) \in \mathbb{R}^p \times \{-1, 1\} \mid i = 1, 2, \dots, M\}$$
$$= \mathcal{D}_+ \cup \mathcal{D}_-(\mathcal{D}_\pm \coloneqq \{\mathbf{x}_i \in \mathbb{R}^p \mid (\mathbf{x}_i, \pm 1) \in \mathcal{D}\}),$$

where $\mathfrak{L}(\mathbf{x}_i)$ is the binary label assigned to $\mathbf{x}_i$. Since the original Cortes-Vapnik problem was introduced as an NP-hard problem with a hierarchical structure, a simple convex relaxation (the soft margin SVM[1]):

$$\underset{(\mathbf{w},b)\in\mathbb{R}^p\times\mathbb{R}}{\text{minimize}}\ \Psi(\mathbf{w}, b) + C\Phi(\mathbf{w}, b), \qquad (4)$$

where $\Psi(\mathbf{w}, b) \coloneqq \frac{1}{2}\|\mathbf{w}\|^2$,

$$\underbrace{\|\mathbf{w}\|\left[\sum_{\mathbf{z}^+\in\mathcal{D}_+} \mathrm{d}\left(\mathbf{z}^+, \Pi_{(\mathbf{w},b)}^{\geq+1}\right) + \sum_{\mathbf{z}^-\in\mathcal{D}_-} \mathrm{d}\left(\mathbf{z}^-, \Pi_{(\mathbf{w},b)}^{\leq-1}\right)\right]}_{=:\Phi(\mathbf{w},b)}, \text{ and}$$

$\mathrm{d}\left(\cdot, \Pi_{(\mathbf{w},b)}^{\geq+1}\right)$ and $\mathrm{d}\left(\cdot, \Pi_{(\mathbf{w},b)}^{\leq-1}\right)$ respectively stand for distances to closed half-spaces $\Pi_{(\mathbf{w},b)}^{\geq+1} \coloneqq \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top\mathbf{x} - b \geq +1\}$ and $\Pi_{(\mathbf{w},b)}^{\geq-1} \coloneqq \{\mathbf{x} \in \mathbb{R}^p \mid \mathbf{w}^\top\mathbf{x} - b \leq -1\}$,

has been used extensively with a *tuning parameter* $C > 0$. However this naive relaxation induces a natural question: **Is the solution of (4) for general training data $\mathcal{D}$ really a mathematically sound extension of the classical SVM ?** This is because, by the complete loss of the hierarchical structure, the solution of (4) cannot reproduce, in general, the classical SVM that maximizes the margin among all error-free linear classifiers for linearly separable training dataset (Note: The above question is common even for the soft-margin SVM applied to the transformed data with nonlinear kernels [1]). Therefore, we newly formulate

$$\text{minimize } \Psi(\mathbf{w}^\star, b^\star) \text{ subject to}$$
$$(\mathbf{w}^\star, b^\star) \in \mathcal{S}_\star \coloneqq \underset{(\mathbf{w},b)\in\mathbb{R}^p\times\mathbb{R}}{\operatorname{argmin}}\ \Phi(\mathbf{w}, b) \qquad (5)$$

as a much more faithful convex relaxation of the original Cortes-Vapnik problem than (4). Remark that the hierarchical convex relaxation (5) is well-defined even for linearly non-separable training dataset $\mathcal{D}$ and can reproduce perfectly the classical SVM for linearly separable dataset unlike (4). Fortunately, the problem (5) falls in the class of the hierarchical convex optimization problems of type (1) and (2), and therefore is solvable efficiently by combining the ideas in the hybrid steepest descent method and the art of proximal splitting [13].

[1] The function $\Phi(\mathbf{w}, \mathrm{b})$ is often expressed with a *hinge loss function* and serves as a convex relaxation of the number of misclassified training samples. The squared margin of the linear classifier with $(\mathbf{w}, \mathrm{b})$ is given by $2^{-1}\Psi^{-1}(\mathbf{w}, \mathrm{b})$.

References

[1] V. N. Vapnik, Statistical Learning Theory, Wiley, 1998.

[2] Y. A. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, 521, pp. 436-444, 2015.

[3] S. Theodoridis, K. Slavakis, and I. Yamada "Adaptive learning in a world of projections: a unifying framework for linear and nonlinear classification and regression tasks," IEEE Signal Process. Mag., 21, pp. 97-123, 2011.

[4] H. Poincaré, Science et Méthode, 1908 (translated by F. Maitland in 1952 as *Science and Method* which is collected in The Value of Science - Essential Writings of Henri Poincaré , pp. 357-572, Random House, 2001).

[5] Y. Habu and H. Shinohara, The present and future of AI, http://www.ntt.co.jp/activity/en/innovation/habu/, 2017.

[6] S. Freud, New Introductory Lectures on Psycho-Analysis (The complete psychological works of S. Freud, J. Strachey.ed.) W. W. Norton and Company, 1990.

[7] S. Shimojo, Saburiminaru inpakuto (in Japanese), Chikuma Shobo, Inc., 2008.

[8] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proc. Natl. Acad. Sci. U.S.A., 79, pp.2554-2558, 1982.

[9] M. E. Raichle, "The brain's dark energy," Scientific American, pp.44-49, March 2010.

[10] A. N. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," Soviet Math. Dokl., 4, pp.1035-1038, 1963.

[11] H. H. Bauschke and P. L. Combettes, Convex Analysis and Monotone Operator Theory in Hilbert Space 2nd ed., Springer, 2017.

[12] P. L. Combettes and I. Yamada, "Compositions and convex combinations of averaged nonexpansive operators," J. Math. Anal. Appl., 425, pp. 55–70, 2015.

[13] I. Yamada and M. Yamagishi, "Hierarchical Convex Optimization by the hybrid steepest descent method with proximal splitting operators - Enhancements of SVM and Lasso," In: H. H. Bauschke, D. R. Luke, R. Burachik eds., Proceedings of Splitting Algorithms, Modern Operator Theory, and Applications 2017, 74 pp, Springer (in press).

[14] I. Yamada, "The hybrid steepest descent method for the variational inequality problem over the intersection of fixed point sets of nonexpansive mappings," In: D. Butnariu, Y. Censor and S. Reich, eds., Inherently Parallel Algorithm for Feasibility and Optimization and Their Applications, Elsevier, pp.473-504, 2001.

[15] C. Cortes and V. N. Vapnik, "Support Vector Networks," Machine Learning, 20, pp.273-297, 1995.