NIPS 2016 debriefing

AOC project-team

The NIPS conference this year was huge, with around 5000 participants, 2500 submissions, and 568 accepted papers of which 46 were selected as oral presentations. The topics that received much attention include:

- Deep Learning: a lot of papers about algorithms and applications, but also a few very recent theoretical results.
- Generative Adversarial Networks: an emerging topic that Yann LeCun described as *the most interesting idea in the last 10 years in ML, in [his] opinion.* [1]
- Optimization: a lot of papers on convex but also nonconvex optimization—in connection with applications, e.g., in matrix factorization or deep learning.
- Clustering: several papers on the theoretical properties of the EM algorithm in Gaussian mixture models.
- Online learning: quite a lot of papers on bandits or related online learning topics.
- High-dimensional learning with structure: works on learning with high-dimensional sparse data, with applications, e.g., in image or audio processing.

We provide below a very selective—thus not comprehensive—list of papers that were presented during the conference or the subsequent workshops.

1 Online learning

Bandits In total more than 20 papers were about bandit optimization. Various problems were addressed: contextual bandits, combinatorial bandits, dueling bandits, adversarial bandits, convex bandits, Gaussian process bandit optimization, as well as new variants such as fair bandits and causal bandits. Below are a few examples:

- Theoretical developments within the classical MAB problem:
  - Garivier et al. (2016) prove that, in the stochastic two-armed bandit problem with Gaussian rewards, strategies based on an exploration phase followed by an exploitation phase are necessarily suboptimal. Thus optimal strategies must mix exploration and exploitation.
  - In the adversarial MAB setting, Gerchinovitz and Lattimore (2016) derive lower bounds showing that existing *refined* upper bounds (i.e., regret bounds in high probability or of the first or second-order type) are essentially optimal.

[1]See `https://goo.gl/JK2Z2f`.

- New variants of the bandit problem:
  - Joseph et al. (2016) introduce the notion of *fairness* in bandits, which forces bandit algorithms to behave with high probability in a way such that, for every pair of arms (i_1, i_2), arm i_1 is played more often than arm i_2 only if arm i_1 has a larger (unknown) mean reward than i_2. This fairness constraint has natural applications in practice. From a theoretical viewpoint, the authors show an interesting gap in the achievable regret guarantees between fair and unfair bandit algorithms.
  - Lattimore et al. (2016) introduce the problem of *causal bandits*. In this setting the goal is to learn the best *intervention* in a known causal model among a set of fixed interventions. The authors provide an algorithm along with regret guarantees that improve on the performances that we would obtain treating interventions as arms while ignoring the underlying causal graph.
  - Kandasamy et al. (2016b) formalize the notion of a multi-armed bandit problem with *multi-fidelity* observations. That is, at every round, the learner picks an arm and can then choose to observe an approximate reward with one among various possible approximation levels (the best approximation is expensive, while a low approximation is cheap). The authors provide an algorithm along with regret guarantees, where the regret is a combination of the usual notion (difference of arms' expectations) and of the costs associated to the approximation levels used by the learner. In another related paper Kandasamy et al. (2016a) adapt this algorithm to the Gaussian Process bandit optimization problem with multi-fidelity observations.

Adaptive online learning Algorithms designed for adversarial (worst-case) online learning are robust, yet they may be criticized for their very unambitious goal. We can indeed usually expect practical performances to be much better than those possible in the worst case, because datasets are often not generated in an adversarial manner. Recently some papers have addressed the *best of both worlds* trade-off: the idea is to design algorithms with optimal worst-case guarantees yet much better performances when used on easier data (e.g., i.i.d. data with a nice distribution). At least two papers in NIPS this year were in this spirit:

- van Erven and Koolen (2016) design an adaptive algorithm for online convex optimization, with regret guarantees of the form

$$\forall u \in \mathcal{U}, \quad \sum_{t=1}^{T} f_t(w_t) - \sum_{t=1}^{T} f_t(u) \lesssim \min\left\{ \sqrt{V_T^u d \ln T} + d \ln T,\ \sqrt{T \ln \ln T} \right\} ,$$

  where $\mathcal{U} \subseteq \mathbb{R}^d$ is a closed convex set, where $f_t : \mathcal{U} \to \mathbb{R}$ and $w_t \in \mathcal{U}$ are the loss function and the point played by the algorithm at time t, and where $V_T^u = \sum_{t=1}^{T} \big((u - w_t)^T g_t \big)^2$ is a variance term that can be sublinear with T under "nice" settings.

- Koolen et al. (2016) show that the above regret guarantee implies fast rate (faster that $\sqrt{T}$) when provided with i.i.d. data under some margin assumption on the underlying distribution.

2 Optimization

2.1 Stochastic methods and finite sums

This topic represented an important part of optimization–related contributions at the confenrence, starting with the tutorial session of Francis Bach and Suvrit Sra: "Large-Scale Optimization: Beyond Stochastic Gradient Descent and Convexity". Papers presented at the main conference include:

- New methods (Defazio, 2016) and new analysis of variants (Shamir, 2016).
- Extensions, to the Riemannian setting (Zhang et al., 2016), to the saddle point setting (Palaniappan and Bach, 2016).
- Lower complexity bounds (Arjevani and Shamir, 2016; Woodworth and Srebro, 2016).

The topic was also very present in the subsequent "Optimization for Machine Learning" workshop with constributions including:

- Oracle Complexity of Second-Order Methods for Finite-Sum Problems (invited oral). Ohad Shamir, Yossi Arjevani.
- Finite Sum Acceleration vs. Adaptive Learning Rates for the Training of Kernel Machines on a Budget. Tobias Glasmachers.
- Harder, Better, Faster, Stronger Convergence Rates for Least-Squares Regression. Aymeric Dieuleveut, Nicolas Flammarion, Francis Bach.
- Asaga: Asynchronous Parallel SAGA. Rémi Leblond, Fabian Pedregosa, Simon Lacoste-Julien.
- Riemannian stochastic variance reduced gradient on Grassmann manifold. Hiroyuki Kasai, Hiroyuki Sato, Bamdev Mishra.
- Stochastic Optimization with Variance Reduction for Infinite Datasets with Finite Sum Structure. Alberto Bietti, Julien Mairal.
- SVRG++ with Non-uniform Sampling. Tamas Kern, Andras Gyorgy.
- Multiple Kernel Learning via Multi-Epochs SVRG. Mitchel Alioscha-Perez, Meshia C?dric Oveneke, Dongmei Jiang, Hichem Sahli.

2.2 Global analysis of non convex problems

A significant amount of optimization–related material presented at the conference focused on the analysis of very specific (sometimes simplified) optimization problems arising from machine learning applications. Typical results imply that the geometry of the specific type of functional landscape considered is favorable despite its non convexity, so that local search methods should perform better than in a worst–case scenarios (e.g. converge to or approximate global solutions). Such conference papers include.

- Matrix completion: Ge et al. (2016); low rank recovery: Bhojanapalli et al. (2016).
- Deep learning: Kawaguchi (2016).
- EM algorithm for a mixture of two gaussians (with known variance): Xu et al. (2016).

- Semi definite programming: Boumal et al. (2016).

The topic was also well represented in the subsequent workshops.

- Globally Optimal Structured Low-Rank Matrix Factorizations. Rene Vidal, invited oral at the Learning in High Dimensions with Structure workshop, see also Haeffele and Vidal (2016).
- Semidefinite Programs with a Dash of Smoothness. Nicolas Boumal, invited presentation at the Optimization for ML workshop.
- Taming non-convexity via geometry in Nonconvex Optimization for Machine Learning: Theory and Practice. Suvrit Sra, presentation at the nonconvex Optimization workshop.
- Gradient descent efficiently finds the cubic-regularized non-convex Newton step. Yair Carmon and John Duchi, Nonconvex Optimization Workshop.
- Understanding the Landscape of Over-complete Tensors Decomposition. Tengyu Ma and Rong Ge, Nonconvex Optimization workshop.

References

Y. Arjevani and O. Shamir. Dimension-free iteration complexity of finite sum optimization problems. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems 29*, 2016. URL http://papers.nips.cc/paper/6271-global-optimality-of-local-search-for-low-rank-matrix-recovery.pdf.

Nicolas Boumal, Vlad Voroninski, and Afonso Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems 29*, 2016. URL http://papers.nips.cc/paper/6517-the-non-convex-burer-monteiro-approach-works-on-smooth-semidefinite-programs.pdf.

A. Defazio. A simple practical accelerated method for finite sums. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

A. Garivier, T. Lattimore, and E. Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016. URL https://papers.nips.cc/paper/6179-on-explore-then-commit-strategies.

R. Ge, J.D. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016. URL http://papers.nips.cc/paper/6048-matrix-completion-has-no-spurious-local-minimum.

S. Gerchinovitz and T. Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016. URL https://papers.nips.cc/paper/6277-refined-lower-bounds-for-adversarial-bandits.

B. Haeffele and R. Vidal. Global optimality in tensor factorization, deep learning, and beyond. 2016. URL https://arxiv.org/abs/1506.07540.

M. Joseph, M. Kearns, H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016. URL https://papers.nips.cc/paper/6355-fairness-in-learning-classic-and-contextual-bandits.

K. Kandasamy, G. Dasarathy, B. Oliva, J. Schneider, and B. Poczos. Gaussian process bandit optimisation with multi-fidelity evaluations. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016a. URL `https://papers.nips.cc/paper/6118-gaussian-process-bandit-optimisation-with-multi-fidelity-evaluations`.

K. Kandasamy, G. Dasarathy, B. Poczos, and J. Schneider. The multi-fidelity multi-armed bandit. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016b. URL `https://papers.nips.cc/paper/6592-the-multi-fidelity-multi-armed-bandit`.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems 29*, 2016. URL `http://papers.nips.cc/paper/6112-deep-learning-without-poor-local-minima.pdf`.

W. M. Koolen, P. Grünwald, and T. van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016. URL `https://papers.nips.cc/paper/6474-combining-adversarial-guarantees-and-stochastic-fast-rates-in-online-learning`.

F. Lattimore, T. Lattimore, and M.D. Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016. URL `https://papers.nips.cc/paper/6195-causal-bandits-learning-good-interventions-via-causal-inference`.

B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

Ohad Shamir. Without-replacement sampling for stochastic gradient methods. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

T. van Erven and W. M. Koolen. Metagrad: multiple learning rates in online learning. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016. URL `https://papers.nips.cc/paper/6268-metagrad-multiple-learning-rates-in-online-learning`.

B. Woodworth and N. Srebro. Tight complexity bounds for optimizing composite objectives. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.

Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems 29*, 2016. URL `http://papers.nips.cc/paper/6047-global-analysis-of-expectation-maximization-for-mixtures-of-two-gaussians.pdf`.

H. Zhang, S. Reddi, and S. Sra. Fast stochastic optimization on riemannian manifolds. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016.