Exploring the intersection of active learning and stochastic convex optimization

Aaditya Ramdas, Aarti Singh
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA - 15213
aramdas, aarti@cs.cmu.edu

***Abstract*—First order stochastic convex optimization is an extremely well-studied area with a rich history of over a century of optimization research. Active learning is a relatively newer discipline that grew independently of the former, gaining popularity in the learning community over the last few decades due to its promising improvements over passive learning. Over the last year, we have uncovered concrete theoretical and algorithmic connections between these two fields, due to their inherently sequential nature and decision-making based on feedback of earlier choices, that have yielded new methods and proofs techniques in both fields. In this note, we lay down the foundations of these connections and summarize our recent advances.**

***Index Terms*—stochastic convex optimization, active learning, tsybakov noise condition, uniform convexity, optimal adaptive algorithms**

I. First-Order Stochastic Convex Optimization

Consider an unknown function f on a bounded set $S \subset \mathbb{R}^d$, with minimizer $x^* = \arg\min_{x \in S} f(x)$ that is k-uniformly convex (k-UC) and L-Lipschitz, i.e. for constants $L, \lambda > 0, k \geq 2$, we have for all $x, y \in S$,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \lambda \|x - y\|_2^k \qquad (1)$$

$$|f(x) - f(y)| \leq L \|x - y\|_2$$

($k = 2$ for strong convexity). A stochastic first order oracle accepts $x \in S$, and returns $\left(\hat{f}(x), \hat{g}(x)\right)$ which are unbiased i.e. $\mathbb{E}\big[\hat{f}(x)\big] = f(x)$, $\mathbb{E}\big[\hat{g}(x)\big] \in \partial f(x)$, and have bounded variance. An optimization algorithm sequentially queries an oracle at points in S and returns $\hat{x}_T$ as an estimate of the optimum of f after T queries (or alternatively tries to achieve a target error of ϵ). Its performance can be measured by function error $\rho_T^* := f(\hat{x}_T) - f(x^*)$ or point error $\epsilon_T^* := \|\hat{x}_T - x^*\|_2$.

II. Active One-Dimensional Threshold Learning

We deal with a bounded interval $S \subset \mathbb{R}$, where every point $x \in S$ has a label $y \in \{+, -\}$ that is drawn from a continuous unknown conditional distribution $\eta(x) := \Pr\left(Y = + | X = x\right)$ that has a unique point t with $\eta(x) = 1/2$. It is common to characterize the slope of the regression function $\eta(x)$ around threshold t, as given by Tsybakov's Noise Condition (TNC)

$$M|x - t|^{k-1} \geq |\eta(x) - \tfrac{1}{2}| \geq \mu |x - t|^{k-1} \text{ if } |\eta(x) - \tfrac{1}{2}| \leq \epsilon_0$$

for some constants $M > \mu > 0, \epsilon_0 > 0, k \geq 1$.

The learner sequentially queries T (possibly dependent) points, observing labels drawn from η after each query, with the goal of returning a guess $\hat{x}_T$ as close to t as possible. One can measure accuracy by excess classification risk (expected $0/1$ loss under uniform distribution) of the threshold classifier at $\hat{x}_T$, compared to the Bayes optimal classifier at t, i.e.

$$\mathcal{R}_T^* := \mathrm{Risk}(\hat{x}_T) - \mathrm{Risk}(t) = \int_{\min(\hat{x}_T, t)}^{\max(\hat{x}_T, t)} |2\eta(x) - 1| dx$$

or alternatively by point error $\mathcal{E}_T^* := |\hat{x}_T - t|$.

III. Connections and Results

For $d = 1$, $0 \in \partial f(x^*)$ and Eq. (1) imply $f(x) - f(x^*) \geq \lambda |x - x^*|^k$ and $|g(x)| \geq \lambda |x - x^*|^{k-1}$. Since the oracle is unbiased, x^* is the unique point with $\hat{\eta}(x) := P(\mathrm{sign}(\hat{g}(x)) = +) = \frac{1}{2}$ and we can show $|\hat{\eta}(x) - \frac{1}{2}| \geq |x - x^*|^{k-1}$, i.e. UC implies that the sign of the noisy gradient satisfies TNC. Such intuition carries through for higher dimensional functions and we leverage this to demonstrate the following results.

Minimax rates: Using ideas from active learning, we prove in [1] that the minimax information complexity (ignoring poly-d, poly-$\log T$ factors) decays similarly with T queries in both fields, specifically $\rho_T^* = \mathcal{R}_T^* = \tilde{\Theta}\left(T^{-\frac{k}{2k-2}}\right)$ and $\epsilon_T^* = \mathcal{E}_T^* = \tilde{\Theta}\left(T^{-\frac{k}{2k-2}}\right)$. Our techniques also yield a $\Omega(T^{-1/2})$ lower bound for all k, for derivative-free (zeroth-order) stochastic optimization which matches known upper bounds.

Adaptivity: In [2], we show that the same strategy can be adopted in both fields to yield algorithms that are adaptive to unknown UC and TNC exponents and constants, achieving the same rates as procedures knowing these parameters.

Sign oracles: Assuming more smoothness (beyond Lipschitz), in [2] we also show that randomized coordinate descent, with efficient line searches using active threshold learning, achieves the same optimal rate for a weak *stochastic sign oracle* that only provides noisy gradient signs in a chosen direction. If the signs are noiseless, it yields exponential convergence rates.

References

[1] A. Ramdas and A. Singh, "Optimal Rates for Stochastic Convex Optimization under Tsybakov Noise Condition", *International Conference on Machine Learning, 2013.*

[2] A. Ramdas and A. Singh, "Algorithmic Connections Between Active Learning and Stochastic Convex Optimization" *submitted.*