Unifying Stochastic Convex Optimization and Active Learning

Aaditya Ramdas, Aarti Singh
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 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. Here, we summarize the foundations of these connections and summarize our recent advances, with special focus on the implications for stochastic optimization. Specifically, we get an interesting lower bound technique that is quite transparent and are simultaneously tight for derivative-free optimization and first-order optimization for both point and function error. We show that a randomized coordinate descent algorithm with an active learning line search can achieve minimax optimal rates while being adaptive to unknown uniform convexity parameters. This procedure only relies on unidirectional noisy gradient signs as opposed to real valued gradient vectors - as a result, rounding errors and other errors that preserve the sign of the gradient lead to deterministic (non-stochastic) rates of convergence.

1 Introduction

This paper summarizes two works from this year [8, 9] that explore the intersection of seemingly distinct fields - convex optimization and active learning. Recently, [7] pointed out their similarity due to their inherent sequential nature and the complex role of feedback in determining future actions. In [8, 9], we make large advances in relating these fields, both intuitively and formally.

At a high level, we show that the role of the regression function in active learning of one-dimensional threshold functions has strong parallels with the sign of a directional gradient in d-dimensional optimization, the function error in optimization plays the role of excess risk in active learning, uniform convexity (of which strong convexity is a special case) is an exact analog of Tsybakov's noise/margin condition in active learning, and so on.

As we shall summarize later in this paper, this leads to explicit mathematical relationships between the two fields - the minimax point error scales the same way in both settings, the function error and excess risk have the same decay rate, lower bound techniques for optimization can be borrowed from active learning, upper bound ideas from optimization yield new methods in active learning, active learning can be used for line-search in descent procedures - the list is long and interesting.

We will first concretely introduce the settings for optimization and learning, and then mathematically summarize all our contributions that include new lower and upper bounds for optimization.

1.1 Setup of First-Order Stochastic Optimization of Uniformly Convex Functions

First-order (or zeroth-order) stochastic convex optimization [6] is the task of approximately minimizing a convex function over a convex set, given oracle access to unbiased estimates of the function and gradient (or just function for zeroth-order) at any point, using as few queries as possible.

Assume we are given an arbitrary convex set $S \subset \mathbb{R}^d$ of known diameter $R = \max_{x,y \in S} \|x-y\|_2$. A convex function f with $x^* = \arg\min_{x \in S} f(x)$ is said to be k-uniformly convex if, for some $\lambda > 0, k \geq 2$ (strong convexity arises when $k = 2$), we have for all $x, y \in S$

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\lambda}{2}\|x - y\|_2^k.$$

f is L-Lipschitz for some $L > 0$ if $\|\nabla f(x)\|_2 \leq L$; equivalently $|f(x) - f(y)| \leq L\|x - y\|_2$ for all $x, y \in S$. A differentiable f is H-strongly smooth (or has a H-Lipschitz gradient) for some $H > \lambda$ if for all $x, y \in S$, we have $\|\nabla f(x) - \nabla f(y)\|_2 \leq H\|x - y\|$, or equivalently

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{H}{2}\|x - y\|_2^2$$

A stochastic first order oracle O is a function that accepts $x \in S$, and returns

$$\left(\widehat{f}(x), \widehat{g}(x)\right) \in \mathbb{R}^{d+1} \text{ where } \mathbb{E}\big[\widehat{f}(x)\big] = f(x), \mathbb{E}\big[\widehat{g}(x)\big] = \nabla f(x)$$

(they also have bounded variance) and the expectation is over any internal randomness of the oracle. An optimization algorithm is a method M that sequentially queries O at points in S and returns $\widehat{x}_T$ as an estimate of the optimum of f after T queries. Denoting $x^*_{f,S} := \arg\min_{x \in S} f(x)$ and $\rho_T(M, f, S, O) := \|\widehat{x}_T - x^*_{f,S}\|_2$, $\epsilon_T(M, f, S, O) := f(\widehat{x}_T) - f(x^*_{f,S})$, we define as in [1]:

$$\epsilon^*_T(\mathcal{F}) := \sup_O \sup_S \inf_M \sup_{f \in \mathcal{F}} \mathbb{E}_O[\epsilon_T(M, f, S, O)] \quad \text{and} \quad \rho^*_T(\mathcal{F}) := \sup_O \sup_S \inf_M \sup_{f \in \mathcal{F}} \mathbb{E}_O[\rho_T(M, f, S, O)]$$

1.2 Setup of Active Threshold Learning under the Tsybakov Noise Condition (TNC)

The problem of one-dimensional threshold estimation assumes you have an interval of length R, say $[0, R]$. Given a point x, it has a label $y \in \{+, -\}$ that is drawn from an unknown conditional distribution (or regression function) $\eta(x) := \Pr\big(Y = +|X = x\big)$ and the threshold t is the unique point where $\eta(x) = 1/2$, with it being larger than half on one side of t and smaller than half on the other (hence it is more likely to draw a $+$ on one side of t and a $-$ on the other side).

The task of active learning of threshold classifiers allows the learner to sequentially query T (possibly dependent) points, observing labels drawn from the unknown conditional distribution after each query, with the goal of returning a guess $\widehat{x}_T$ as close to t as possible. In the formal study of classification (cf. [11]), it is common to study minimax rates when the regression function $\eta(x)$ satisfies Tsybakov's noise or margin condition (TNC) with exponent k at the threshold t.

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

for some constants $M > \mu > 0, \epsilon_0 > 0, k \geq 1$ (this is the version of TNC from [2]).

A standard measure for how well a classifier h performs is given by its risk, which is simply the probability of classification error (expectation under $0-1$ loss), $\mathcal{R}(h) = \Pr\big[h(x) \neq y\big]$. The performance of threshold learning strategies can be measured by the excess classification risk of the resultant threshold classifier at $\widehat{x}_T$ compared to the Bayes optimal classifier at t as given by [2]

$$\mathcal{R}(\widehat{x}_T) - \mathcal{R}(t) = \int_{\widehat{x}_T \wedge t}^{\widehat{x}_T \vee t} |2\eta(x) - 1| dx \tag{2}$$

Akin to [2], we used a uniform marginal distribution for risk of active learning since there is no underlying distribution over x. Alternatively, one can measure the one-dimensional point error $|\widehat{x}_T - t|$ in estimation of the threshold ([2] define minimax rates like the previous subsection).

[1] Note that $|x - t| \leq \delta_0 := \left(\frac{\epsilon_0}{M}\right)^{\frac{1}{k-1}} \implies |\eta(x) - 1/2| \leq \epsilon_0 \implies |x - t| \leq \left(\frac{\epsilon_0}{\mu}\right)^{\frac{1}{k-1}}$

[2] $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$

2 Connecting The TNC and Uniform Convexity Exponents

2.1 Stochastic Gradient-Sign Oracles

Define a stochastic sign oracle to be a function of $x \in S, j \in \{1...d\}$, that returns $\widehat{s}_j(x) \in \{+,-\}$ as a noisy sign$\big([\nabla f(x)]_j\big)$, where $[\nabla f(x)]_j$ is the j-th coordinate of ∇f and [3]

$$\big|\eta_j(x) - 0.5\big| = \Theta\Big(\big|[\nabla f(x)]_j\big|\Big) \text{ and } \eta_j(x) = \Pr\Big(\widehat{s}_j(x) = \text{sign}\big([\nabla f(x)]_j\big)\big|x\Big)$$

(the probability is over any internal randomness of the oracle).

A sign oracle is weaker than a first order oracle, and can actually be obtained by returning the sign of the first order oracle's noisy gradient if the mass of the noise distribution grows linearly around its zero mean (as satisfied by gaussian, uniform, and many other kinds of noise). At the optimum along coordinate j, the oracle returns a ± 1 with equal probability, and otherwise returns the correct sign with a probability proportional to the value of the directional derivative at that point (this is reflective of the fact that the larger the derivative's absolute value, the easier it would be for the oracle to approximate its sign, hence the smaller the probability of error). It is not unreasonable that there may be other circumstances where even calculating the (real value) gradient in the i-th direction could be expensive, but estimating its sign could be a much easier task as it only requires estimating whether function values are expected to increase or decrease along a coordinate.

2.2 Making Connections Transparent in 1-D

Taking one point as x^* in the definition of UC, we see that $|f(x) - f(x^*)| \geq \frac{\lambda}{2}\|x - x^*\|^k$ Since $\|\nabla f(x)\|\|x - x^*\| \geq \nabla f(x)^\top (x - x^*) \geq f(x) - f(x^*)$ (by convexity),

$$\|\nabla f(x) - 0\| \geq \frac{\lambda}{2}\|x - x^*\|^{k-1} \tag{3}$$

Let us work in 1-D for clarity of exposition, i.e. $|\nabla f(x) - 0| \geq \frac{\lambda}{2}|x - x^*|^{k-1}$. Notice the similarity to the form of the TNC $|\eta(x) - 1/2| \geq \mu|x - t|^{k-1}$.

Since f is convex, its noiseless gradient is an increasing function of x that is negative before x^* and positive after x^*. Hence, sign$(\nabla f(x))$ is the true label of x, sign$(\nabla f(x) + z)$ is the observed label, and x^* corresponds to the decision boundary where labels switch signs. Defining

$$\eta_f(x) := \Pr\Big(\text{sign}(\nabla f(x) + z) = +|x\Big)$$

then minimizing f corresponds to identifying the Bayes classifier $[x^*, 1]$ because the point at which $\eta_f(x) = 0.5$ is where $\nabla f(x) = 0$, which is x^*.

Consider a point x with $\nabla f(x) > 0$ and hence has true label $+$. The probability of seeing a $+$ is the probability that we draw $z \in \big(-\nabla f(x), \infty\big)$ so the sign of $\nabla f(x) + z$ is still positive. Hence,

$$\begin{aligned}
\eta_f(x) \;&=\; \Pr\Big(\nabla f(x) + z > 0\Big) \\
&=\; \Pr(z > 0) + \Pr\Big(-\nabla f(x) < z < 0\Big) \;=\; 0.5 + \Theta\Big(\nabla f(x)\Big) \\
&\implies \big|\eta_f(x) - \tfrac{1}{2}\big| \;=\; \Theta\Big(\nabla f(x)\Big) \;=\; \Theta\Big(|x - x^*|^{k-1}\Big)
\end{aligned}$$

Hence, $\eta_f(x)$ satisfies the TNC with exponent k, and an active learning algorithm can be used to obtain a point $\widehat{x}_T$ with small point-error and excess risk. Note that function error in convex optimization is bounded above by excess risk of the corresponding active learner using eq (2) because

$$\begin{aligned}
f_j(\widehat{x}_T) - f_j(x_j^*) \;&=\; \left|\int_{\widehat{x}_T \wedge x_j^*}^{\widehat{x}_T \vee x_j^*} [\nabla f(x)]_j \mathrm{dx}\right| \qquad = \Theta\left(\int_{\widehat{x}_T \wedge x_j^*}^{\widehat{x}_T \vee x_j^*} |2\eta_f(x) - 1| \mathrm{dx}\right) \\
&=\; \Theta\Big(\mathcal{R}(\widehat{x}_T)\Big)
\end{aligned}$$

[3] $f = \Theta(g)$ means $f = \Omega(g)$ and $f = \mathrm{O}(g)$ (rate of growth)

3 Summary of Contributions

3.1 Lower Bounds

Let $\mathcal{F}^{SC}$ be the set of all Lipschitz strongly convex functions, and $\mathcal{F}^{C}$ the set of all Lipschitz convex functions. Let $\mathcal{F}_k^{UC}$ be the set of all Lipschitz uniformly convex functions with exponent $k \geq 2$. Let $\mathcal{F}_k$ be the set of all functions that satisfy $f(x) - f(x^*) \geq c\|x - x^*\|_2^k$ for some $c > 0, k \geq 1$. This forms a nested hierarchy of classes of $\mathcal{F}^C$, with $\mathcal{F}_{k_1} \subset \mathcal{F}_{k_2}$ whenever $k_1 < k_2$. Also notice that $\mathcal{F}_2 \supseteq \mathcal{F}^{SC} = \mathcal{F}_2^{UC}$ and $\bigcup_k \mathcal{F}_k \subseteq \mathcal{F}^C$. For any finite $1 \leq k < \infty$, the function is strictly convex and hence the minimizer is well-defined and unique.

Theorem 1. *For first order oracles, we have* $\epsilon_T^*(\mathcal{F}_k) = \Theta(T^{-\frac{k}{2k-2}})$ *and* $\rho_T^*(\mathcal{F}_k) = \Theta(T^{-\frac{1}{2k-2}})$.

Since we use uniformly convex functions in our proofs, the bounds immediately hold for uniformly convex functions too. Also, since $\bigcup_k \mathcal{F}_k \subseteq \mathcal{F}^C$, these bounds are valid for $\mathcal{F}^C$ too (specifically taking $\kappa \to \infty$). These lower bounds (which are tight) have *exactly* the same rate as the lower bounds for point-error and risk in active learning [2], emphasizing our claimed connections.

Theorem 2. *For zeroth order oracles, we have* $\epsilon_T^*(\mathcal{F}_k) = \Omega(1/\sqrt{T})$ *and* $\rho_T^*(\mathcal{F}_k) = \Omega(T^{-\frac{1}{2k}})$.

Some of these bounds were known - strongly convex functions in [1] (first order), uniformly convex functions in [10] (first order, completely different proof technique) and strongly convex functions in [5] (zeroth order), while they were not published for zeroth order uniformly convex functions. In a single simple unified proof, we simultaneously get all these lower bounds, using techniques from the active learning literature (specifically, the proofs in [2]). We also simultaneously get lower bounds for function and point error, while most of the literature has focused only on function error.

3.2 Upper Bounds

The first algorithm is a generalization of Epoch Gradient Descent [3] that minimized strongly convex functions at the optimum rate. However, this needs knowledge of Lipschitz constant L, uniform convexity constant λ and exponent k. This procedure works for $\mathcal{F}_k$ for $k \geq 1$, which actually captures more functions than $\mathcal{F}_k^{UC}$ which only holds for $k \geq 2$. In fact as $k \to 1$, we get exponentially fast rates just as in the case of active threshold learning [2].

Theorem 3. *Algorithm* $EpochGD(S, \kappa, T, \delta, L, \lambda)$ *[8] returns* $\widehat{x}_T \in S$ *after* T *queries to any stochastic first order oracle* O*, such that for any* $f \in \mathcal{F}^\kappa, \kappa > 1$ *on any* $S \in \mathbb{S}$, $f(\widehat{x}_T) - f(x_f^*) = \widetilde{O}(T^{-\frac{\kappa}{2\kappa-2}})$ *and* $\|\widehat{x}_T - x_f^*\| = \widetilde{O}(T^{-\frac{1}{2\kappa-2}})$ *hold with probability at least* $1 - \delta$ *for any* $\delta > 0$. [4]

The second algorithm is more general because it works with the much weaker stochastic sign oracle, and is adaptive to all unknown uniform convexity and smoothness parameters. This is very powerful but it comes at the price of requiring a stronger smoothness Assumption LkSS [9] which is analogous to a two-sided condition in Eq. (3), very much like the two-sided TNC condition in Eq. (1). First, we use ideas from [4] and exploit connections between the two areas to develop an optimal active threshold learning algorithm that is adaptive to all unknown TNC parameters.

Theorem 4. *In the setting of one-dimensional active learning of thresholds, Algorithm 1 in [9] adaptively achieves* $\mathcal{R}(x_E) - \mathcal{R}(t) = \tilde{\mathrm{O}}\left(T^{-\frac{k}{2k-2}}\right)$ *with probability at least* $1 - \delta$ *in* T *queries when the unknown regression function* $\eta(x)$ *has unknown TNC parameters* μ, k.

We then repeatedly use this adaptive 1-D active learner to perform line-search in a randomized coordinate descent method to get an optimization algorithm that is adaptive to unknown parameters. As a special case, strongly convex and strongly smooth functions can be minimized in $\tilde{\mathrm{O}}(1/T)$ steps.

Theorem 5. *Given access to only a stochastic sign oracle, Randomized Stochastic-Sign Coordinate Descent [9] can minimize UC and LkSS functions at the minimax optimal convergence rate for expected function error of* $\tilde{\mathrm{O}}(T^{-\frac{k}{2k-2}})$ *adaptive to all unknown convexity and smoothness parameters.*

[4] $\widetilde{O}$ hides $\log\log T$ and $\log(1/\delta)$ factors

References

[1] A. Agarwal, P.L. Bartlett, P. Ravikumar, and M.J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *Information Theory, IEEE Transactions on*, 58(5):3235–3249, 2012.

[2] R.M. Castro and R.D. Nowak. Minimax bounds for active learning. In *Proceedings of the 20th annual conference on learning theory*, pages 5–19. Springer-Verlag, 2007.

[3] E. Hazan and S. Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 23nd Annual Conference on Learning Theory*, 2011.

[4] A. Iouditski and Y. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. *Universite Joseph Fourier, Grenoble, France*, 2010.

[5] K.G. Jamieson, R.D. Nowak, and B. Recht. Query complexity of derivative-free optimization. *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[6] A.S. Nemirovski and D.B. Yudin. *Problem complexity and method efficiency in optimization.* John Wiley & Sons, 1983.

[7] M. Raginsky and A. Rakhlin. Information complexity of black-box convex optimization: A new look via feedback information theory. In *47th Annual Allerton Conference on Communication, Control, and Computing, 2009.*, 2009.

[8] A. Ramdas and A. Singh. Optimal rates for stochastic convex optimization under tsybakov noise condition. *Intl. Conference in Machine Learning (ICML)*, 2013.

[9] A. Ramdas and A. Singh. Algorithmic connections between active learning and stochastic convex optimization. *Algorithmic Learning Theory (ALT)*, 2013.

[10] K. Sridharan and A. Tewari. Convex games in banach spaces. In *Proceedings of the 23nd Annual Conference on Learning Theory*, 2010.

[11] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.