Affiliations:

Carnegie Mellon University

MACHINE LEARNING DEPARTMENT

Exploring the Intersection of Active Learning and Stochastic Convex Optimization

Aaditya Ramdas and Aarti Singh

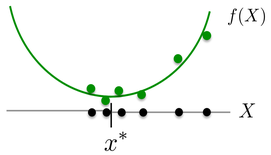
Key Papers (by Ramdas, Singh) :

Optimal Rates for Stochastic Convex Optimization under Tsybakov Noise Condition (ICML '13)

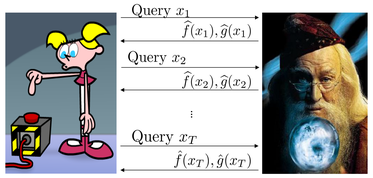
Algorithmic Connections between Active Learning and Stochastic Convex Optimization (ALT '13)

Introduction

First Order d-D Stochastic Convex Optimization



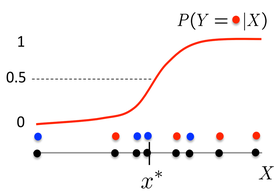
Minimize # queries needed to find optimum (information complexity)



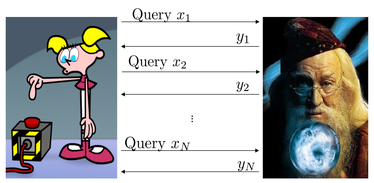
$\mathbb{E}[\hat{f}(x)] = f(x) \in \mathbb{R}$ and $\mathbb{E}[\hat{g}(x)] \in \partial f(x) \subset \mathbb{R}^d$ with variance σ^2

Point error: $\|x_T - x^*\|$, Function error: $f(x_T) - f(x^*)$

Active 1-D Threshold Learning



Minimize # queries needed to find decision boundary (sample complexity)



$y_i \in \{0, 1\}$ and $\mathbb{E}[Y|X = x] = P(Y = 1|X = x)$

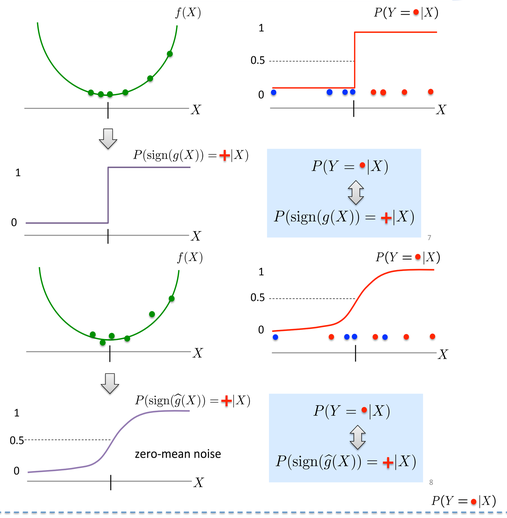
Point error: $|x_N - x^*|$, Excess risk: $\text{Risk}(x_N) - \text{Risk}(x^*)$

References:

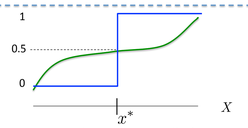
Information theoretic lower bounds on the oracle complexity of stochastic convex optimization (Agarwal, Bartlett, Ravikumar, Wainwright, 2010)

Information complexity of black-box convex optimization (Raginsky, Rakhlin, 2009)

Key Ideas



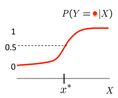
Active Learning and TNC



- JUMP : Binary search - exponentially fast! $\kappa = 1$
- FLAT : No intelligent queries – passive learning $\kappa \to \infty$

If Tsybakov's Noise Condition (TNC) holds

$$|P(Y = \bullet|X = x) - 1/2| \geq \lambda \|x - x^*\|^{\kappa - 1}$$



then minimax optimal active learning rate in 1-dim is

$$\mathbb{E}[\|\widehat{x}_N - x^*\|] \asymp N^{-\frac{1}{2\kappa-2}}$$

and under 0/1 loss + smoothness of P(Y|X)

$$\text{Risk}(\widehat{x}_N) - \text{Risk}(x^*) \asymp N^{-\frac{\kappa}{2\kappa-2}}$$

$N^{-\frac{1}{2}}$, N^{-1} $\Big\}$ $\kappa = 2$

Uniform Convexity and TNC

- Strong convexity $\equiv$ TNC with $\kappa = 2$

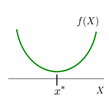
$$f(y) \geq f(x) + g(x)^\top (y - x) + \lambda \|y - x\|^2$$
$$\Rightarrow \quad f(x) - f(x^*) \geq \lambda \|x - x^*\|^2$$
$$\Rightarrow \quad \|g(x) - \cancel{g(x^*)}^{0}\| \geq \lambda \|x - x^*\|$$

- If noise pdf grows linearly around its zero mean (Gaussian, uniform, etc), then (for a different constant)

$$|P(\text{sign}(\hat{g}(X)) = +|X = x) - 1/2| \geq \lambda \|x - x^*\|$$

- TNC for convex functions $\kappa \geq 1$

$$f(x) - f(x^*) \geq \lambda \|x - x^*\|^\kappa$$
$$\Rightarrow \quad \|g(x) - g(x^*)\| \geq \lambda \|x - x^*\|^{\kappa - 1}$$



Controls strength of convexity around the minimum

Uniformly convex (UC) function implies TNC $\kappa \geq 2$

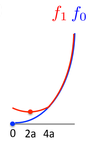
$$f(y) \geq f(x) + g(x)^\top (y - x) + \lambda \|y - x\|^k$$

Controls strength of convexity everywhere in domain

Minimax Lower Bounds

Technique From Active Learning

$$\sup_O \sup_S \inf_{\widehat{x}} \sup_f \mathbb{E}[\|\widehat{x} - x_f^*\|] = \Omega(T^{-\frac{1}{2\kappa-2}})$$



$S^* = [0, 1]^d \cap \{\|x\| \leq 1\}$

$O^* : \ \widehat{f}(x) \sim \mathcal{N}(f(x), \sigma^2), \ \widehat{g}(x) \sim \mathcal{N}(g(x), \sigma^2 \mathbb{I}_d)$

$f_0(x) = c_1 \sum_{i=1}^{d} |x_i|^\kappa$

$$f_1(x) = \begin{cases} c_1(|x_1 - 2a|^\kappa + \sum_{i=2}^{d} |x_i|^\kappa) + c_2 & x_1 \leq 4a \\ f_0(x) & \text{otherwise} \end{cases}$$

$P_0 = P(\{X_i, f_0(X_i), g_0(X_i)\}_{i=1}^T)$ $\quad P_1 = P(\{X_i, f_1(X_i), g_1(X_i)\}_{i=1}^T)$

- Fano's Inequality if $\text{KL}(P_0, P_1) \leq \text{Constant}$

$$\inf_{\widehat{x}} \sup_f P(\|\widehat{x} - x_f^*\| > \|x_{f_0}^* - x_{f_1}^*\|/2) \geq \text{constant}$$

$$KL(P_0, P_1) \leq \frac{T}{2} \left(\max_{x \in [0,1]^d \cap S} \|g_0(x) - g_1(x)\|^2 \right) + \frac{T}{2} \left(\max_{x \in [0,1]^d \cap S} (f_0(x) - f_1(x))^2 \right)$$

Query that yields max difference between function/gradient values

$$= O(Ta^{2\kappa-2}) + O(Ta^{2\kappa})$$
$$\leq \text{Constant} \qquad \boxed{\text{if } \|x_{f_0}^* - x_{f_1}^*\|/2 = a = T^{-\frac{1}{2\kappa-2}}}$$

Yields lower bounds for point+function error and for first +zeroth-order (derivative-free) stochastic optimization.

The Lower Bound Is Tight!

Epoch-Gradient Descent

Initialize $e = 1, x_1^1, T_1, R_1, \eta_1$
until Oracle budget T is exhausted $(E \leq \log T$ epochs $)$
for $t = 1$ to T_e **do**

$$\text{Projected Gradient Descent } x_{t+1}^e = \prod_{S \cap B(x_t^e, R_e)} (x_t^e - \eta_e \hat{g}_t)$$

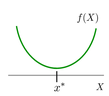
Requires knowledge of κ

$$x_1^{e+1} = \frac{1}{T_e} \sum_{t=1}^{T_e} x_t^e$$
$$T_{e+1} = 2T_e, \eta_{e+1} = \eta_e \cdot 2^{-\frac{\kappa}{2\kappa-2}}, R_{e+1} \sim \eta_{e+1}^{\frac{1}{\kappa}}, \ e \leftarrow e + 1$$

Algorithm's last point achieves minimax rates.

Theorem: The minimax optimal first-order stochastic optimization error rate for d-D Lipschitz convex functions that satisfy

$$f(x) - f(x^*) \geq \lambda \|x - x^*\|^\kappa$$



for some $\kappa > 1$ over a bounded set is

$$\mathbb{E}\left[\|\hat{x}_T - x^*\|\right] = \Theta\left(T^{-\frac{1}{2\kappa-2}}\right)$$
$$\mathbb{E}\left[f(\hat{x}_T) - f(x^*)\right] = \Theta\left(T^{-\frac{\kappa}{2\kappa-2}}\right)$$

$T^{-3/2}$	T^{-1}	$T^{-\frac{1}{2}}$
	Strongly convex	Convex
$\kappa = 3/2$	$\kappa = 2$	$\kappa \to \infty$

Precisely the rates for 1-D active learning!

References:

Minimax bounds for active learning (Castro, Nowak, 2007)

Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization (Hazan, Kale, 2011)

Convex games in banach spaces (Sridharan, Tewari, 2010)

Upper Bounds

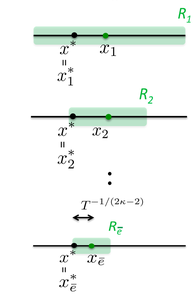
Adaptive 1-D Active Learning

Robust Binary Search

For $e = 1, \ldots, E = \log \sqrt{T/\log T}$ (ignoring κ)
Do passive learning with sample budget $T_e = T/E$
$R_{e+1} = R_e/2, e \leftarrow e + 1$

Adapted from Iouditski-Nesterov'10

$$\exists \bar{e} \text{ s.t. } \|x_{\bar{e}} - x_{\bar{e}}^*\| \preceq T^{-1/(2\kappa-2)}$$
$$x_{\bar{e}}^* = x^*$$
$$\forall e \geq \bar{e}, \ \|x_e - x_{\bar{e}}\| \preceq T^{-1/(2\kappa-2)}$$



Initially, our search area is the whole domain.

We keep halving the search radius around the best guess of the previous epoch.

At some epoch we're close to the optimum, which is within our search radius. After that epoch, the shrinking domain ensures we can't stray too far.

Achieves minimax rates, adaptive to unknown exponent, constant!

Adaptive d-D Optimization

Gradient Sign Oracle

Inputs a point and coordinate and returns a noisy 1-bit sign of the gradient's corresponding coordinate such that

$$\Pr\{s_j(x) = \text{sign}([\nabla f(x)]_j)\} \ \propto \ |[\nabla f(x)]_j|$$
$$\text{or} \quad \Pr\{s_j(x) = +\} - 1/2 \ \propto \ [\nabla f(x)]_j$$

Querying near the directional minimum gives uniformly random sign. Far from the minimum, we are likely to get the correct sign.

Randomized Coordinate Descent

Oracle: Stochastic sign oracle returning noisy $\text{sign}([\nabla f(x)]_j)$
BlackBox: Adaptive line search from x in direction d for n steps
Initialize any $x_0 \in S$
for $e = 1$ to E **do** $(E \leq d(\log T)^2$ epochs$)$
Choose $d_e \in \{1 \ldots d\}$ uniformly at random
$x_e \leftarrow$ Adaptive Line Search $(x_{e-1}, d_e, T/E)$
$e \leftarrow e + 1$ Requires no knowledge of κ

If f is "exactly" κ–uniformly convex (like two-sided TNC), then

$$\sup_f \mathbb{E}\left[\|x_1^{E+1} - x^*\|\right] = \tilde{\text{O}}\left(T^{-\frac{1}{2\kappa-2}}\right)$$
$$\sup_f \mathbb{E}\left[f(x_1^{E+1}) - f(x^*)\right] = \tilde{\text{O}}\left(T^{-\frac{\kappa}{2\kappa-2}}\right)$$

Achieves minimax rates, adaptive to unknown exponent, constant!

References:

Primal-dual subgradient methods for minimizing uniformly convex functions (Juditsky, Nesterov, 2009)

Query complexity of derivative-free optimization (Jamieson, Recht, Nowak, 2012)