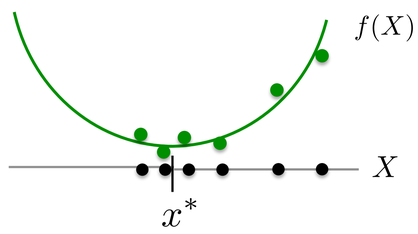
Optimal Rates for Stochastic Convex Optimization under Tsybakov Noise Condition

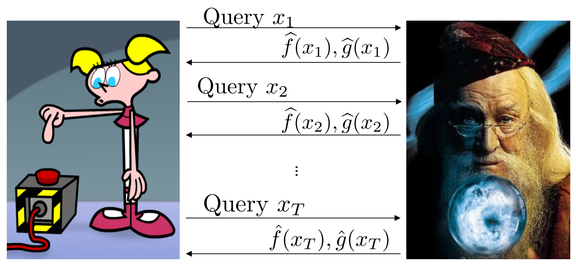
Aaditya Ramdas and Aarti Singh

Introduction

First Order d-D Stochastic Convex Optimization



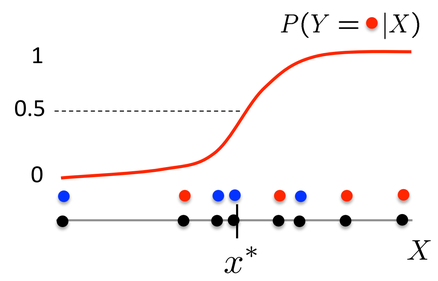
Minimize # queries needed to find optimum (information complexity)



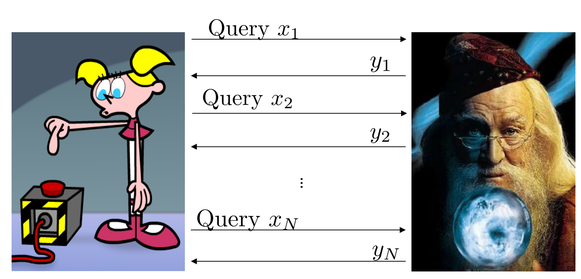
$\mathbb{E}[\widehat{f}(x)] = f(x) \in \mathbb{R}$ and $\mathbb{E}[\widehat{g}(x)] \in \partial f(x) \subset \mathbb{R}^d$ with variance σ^2

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)

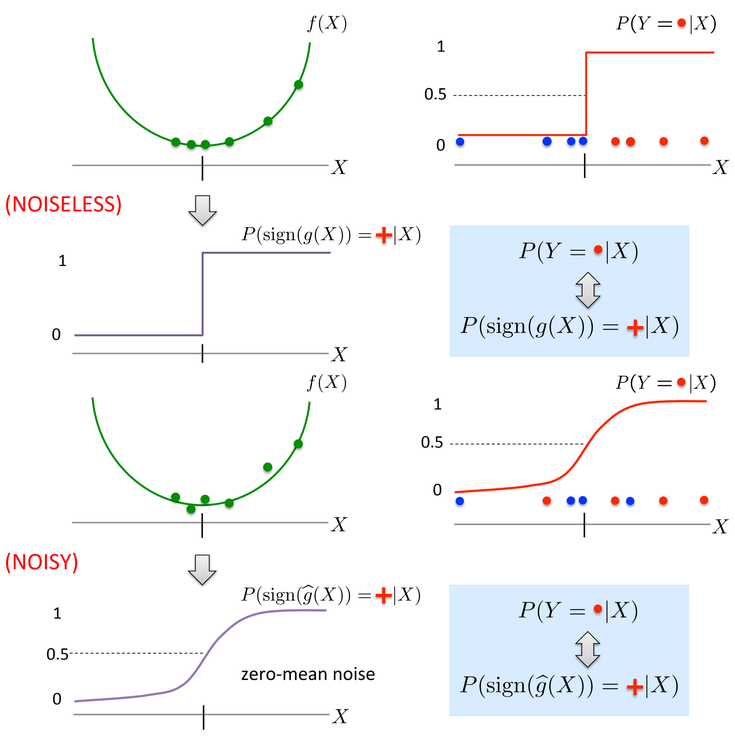


$y_i \in \{0, 1\}$ and $\mathbb{E}[Y|X = x] = P(Y = 1|X = x)$

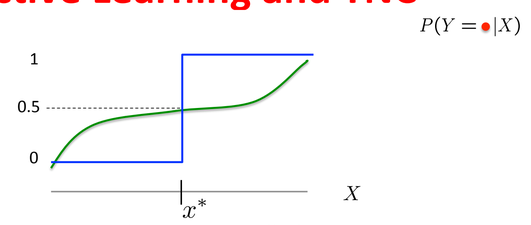
Point error: $|x_N - x^*|$, Excess risk: $\text{Risk}(x_N) - \text{Risk}(x^*)$

UC Implies TNC

Intuition



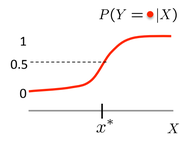
Active Learning and TNC



- 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} } $\kappa = 2$

Uniform Convexity and TNC

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

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

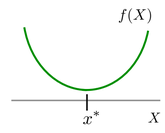
($g(x^*) \to 0$)

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

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

- 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 $\kappa \geq 2$

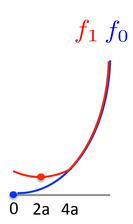
$$f(y) \geq f(x) + g(x)^\top (y - x) + \lambda \|y - x\|^k$$

Controls strength of convexity everywhere in domain

d-D Lower & Upper Bounds

Lower bounds based on 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 $\text{KL}(P_0, P_1) \leq \text{Constant}$

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

$$\text{KL}(P_0, P_1) \leq \frac{T}{2}\left(\max_{x \in S^*} \|g_0(x) - g_1(x)\|^2\right) + \frac{T}{2}\left(\max_{x \in S^*} \|f_0(x) - f_1(x)\|^2\right)$$

Query that yields max difference between function/gradient values

$$= \text{O}\left(Ta^{2\kappa - 2}\right) + \text{O}\left(Ta^{2\kappa}\right)$$

$$\leq \text{Constant} \quad \boxed{\text{if } \|x_{f_0}^* - x_{f_1}^*\|/2 = a = T^{-\frac{1}{2\kappa - 2}}}$$

Immediately yields lower bounds for function error and for zeroth order derivative-free stochastic optimization

Epoch-based Gradient Descent

Initialize $e = 1, x_1^1, T_1, R_1, \eta_1$

until Oracle budget T is exhausted $\quad (E \leq \log T$ epochs $)$

for $t = 1$ to T_e **do**

Projected Gradient Descent $x_{t+1}^e = \prod_{S \cap B(x_1^e, R_e)} (x_t^e - \eta_e \hat{g}_t)$

$x_1^{e+1} = \frac{1}{T_e} \sum_{t=1}^{T_e} x_t^e$ $\quad$ Requires knowledge of κ

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

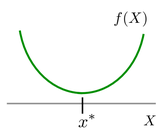
We can show that this algorithm's last point achieves the minimax optimal error rates, giving us the following theorem

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

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

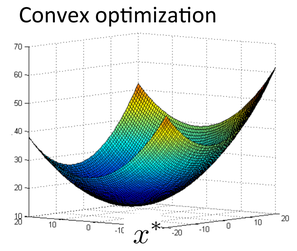


Precisely the rates for 1-D active learning!

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

Takeaway Messages

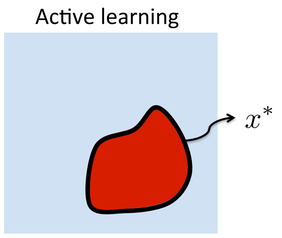
Convex optimization



Minimizer: a point (0-dim)

$T^{-\frac{\kappa}{2\kappa - 2}}$

Active learning



Decision boundary: curve (d-1 dim)

$N^{-\frac{2}{2 + \frac{d-1}{\gamma}}}$

Complexity of convex optimization in any dimension is same as complexity of active learning in 1 dimension.

- There exists a connection between the two fields due to sequential nature and role of feedback
- This can be explicitly captured by the implication of Tsybakov's Noise Condition from Uniform Convexity
- Exploiting this leads to sharing of lower and upper bound ideas, new proof techniques and algorithms

Conclusion

Extensions

- Epoch-based active learning algorithm – adaptive to unknown Tsybakov noise parameters
- Using active learning to perform line search in optimization
- Optimization procedures with access to directional gradient signs (not real valued vectors)
- Randomized coordinate descent algorithm – adaptive to unknown convexity parameters
- New algorithms for derivative-free stochastic optimization?
- Porting active learning procedures to non-convex optimization?

References

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

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

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

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

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

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

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