Optimal convex optimization under Tsybakov noise through connections to active learning

Aarti Singh

Joint work with: Aaditya Ramdas
Connections between convex optimization and active learning (a formal reduction)
Role of feedback

- in convex optimization
  minimize computational complexity
  (# queries needed to find optimum)

- in active learning
  minimize sample complexity
  (# queries needed to find decision boundary)

Raginsky-Rakhlin’09
Active learning oracle model

• Oracle provides $Y \in \{0, 1\}$

\[ \mathbb{E}[Y|X] = P(Y = 1|X) \]
Stochastic optimization oracle model (first-order)

• Oracle provides $f(x), g(x) = \nabla f(x)$

\[
\begin{align*}
\text{Query } x_1 & \quad \rightarrow \quad \hat{f}(x_1), \hat{g}(x_1) \\
\text{Query } x_2 & \quad \rightarrow \quad \hat{f}(x_2), \hat{g}(x_2) \\
& \quad \vdots \\
\text{Query } x_N & \quad \rightarrow \quad \hat{f}(x_N), \hat{g}(x_N)
\end{align*}
\]

• $\mathbb{E}[\hat{f}(x)] = f(x), \mathbb{E}[\hat{g}(x)] = g(x)$ unbiased, variance $\sigma^2$
Connections in 1-dim noiseless setting

• convex optimization

\[ f(X) \]

\[ P(\text{sign}(g(X)) = + | X) \]

• active learning

\[ P(Y = \bullet | X) \]

\[ P(\text{sign}(g(X)) = + | X) \]
Connections in 1-dim noisy setting

- **convex optimization**

\[ f(X) \]

\[ P(\text{sign}(\hat{g}(X)) = + | X) \]

- **active learning**

\[ P(Y = \bullet | X) \]

\[ P(\text{sign}(\hat{g}(X)) = + | X) \]
Minimax active learning rates in 1-dim

- If Tsybakov Noise Condition (TNC) holds
  \[ \kappa \geq 1 \]
  \[ |P(Y = \bullet | X = x) - 1/2| \geq \lambda \|x - x^*\|^{\kappa - 1} \]

then minimax optimal active learning rate in 1-dim is

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

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

\[ \text{Risk}(\hat{x}_N) - \text{Risk}(x^*) \asymp N^{-\frac{\kappa}{2\kappa - 2}} \]

Castro-Nowak’07
TNC and strong convexity

• Strong convexity \( \equiv \) TNC with \( \kappa = 2 \)

\[
f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \lambda \|x - y\|^2
\]

\[
\Rightarrow f(x) - f(x^*) \geq \lambda \|x - x^*\|^2
\]

\[
\Rightarrow \|g(x) - g(x^*)\| \geq \lambda \|x - x^*\|
\]

• If noise pmf grows linearly around its zero mean (Gaussian, uniform, triangular), then

\[
\left| P(\text{sign}(\hat{g}(X)) = \pm | X = x) - 1/2 \right| \geq \lambda \|x - x^*\|
\]
Algorithmic reduction (1-dim)

- In 1-dim, consider any active learning algorithm that is optimal for TNC exponent $\kappa = 2$. When given labels $Y = \text{sign}(\hat{g}(X))$, where $f(x)$ is a strongly convex function with Lipschitz gradients, it yields
  $$\mathbb{E}[\| \hat{x}_T - x^* \|] = O(T^{-\frac{1}{2}})$$
  $$\mathbb{E}[f(\hat{x}_T) - f(x^*)] = O(T^{-1})$$

- Matches optimal rates for strongly convex functions
  Nemirovski-Yudin’83, Agarwal-Bartlett-Ravikumar-Wainwright’10

- What about d-dim?
Complexity of convex optimization in any dimension is same as complexity of active learning in 1 dimension.

Minimizer: a point (0-dim)

Decision boundary: curve (d-1 dim)

$T^{-1}$

$N \leq \frac{2}{2 + \frac{d-1}{\gamma}}$
Algorithmic reduction (d-dim)

Random coordinate descent with 1-dim active learning subroutine

\begin{align*}
\textbf{For } & e = 1, \ldots, E = d(\log T)^2 \\
& \text{Choose coordinate } j \text{ at random from } 1, \ldots, d \\
& \text{Do active learning along coordinate with sample budget } T_e = T/E \\
& \text{treating } \text{sign}(\hat{g}_j(X_t)) \text{ as label } Y_t
\end{align*}

- If \( f \) is strongly convex with Lipschitz gradients

\[
\sup_{\mathcal{O}} \sup_{S} \inf_{\hat{x}} \sup_{f} \mathbb{E}[\|\hat{x} - x^*\|] = \tilde{O}(T^{-\frac{1}{2}})
\]
\[
\sup_{\mathcal{O}} \sup_{S} \inf_{\hat{x}} \sup_{f} \mathbb{E}[f(\hat{x}) - f(x^*)] = \tilde{O}(T^{-1})
\]
Degree of convexity via Tsybakov noise condition (TNC)
Degree of convexity via TNC

• TNC for convex functions \( \kappa \geq 1 \)

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

\[
\Rightarrow \|g(x) - g(x^*)\| \geq \lambda \|x - x^*\|^\kappa^{-1}
\]

Controls strength of convexity around the minimum

• Uniformly convex function implies TNC \( \kappa \geq 2 \)

\[
f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\lambda}{2} \|x - y\|^\kappa
\]

Controls strength of convexity everywhere in domain
Minimax convex optimization rates

**Theorem:** If TNC for convex functions holds $\kappa > 1$

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

and $f$ is Lipschitz, then minimax optimal convex optimization rate over a bounded set (diam $\leq 1$) is

\[ \|\hat{x}_T - x^*\| \asymp T^{-\frac{1}{2\kappa - 2}} \text{ d-dim} \]

\[ \|f(\hat{x}_T) - f(x^*)\| \asymp T^{-\frac{\kappa}{2\kappa - 2}} \text{ d-dim} \]

Precisely the rates for 1-dim active learning!
Lower bounds based on active learning

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

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

\[O^* : \hat{f}(x) \sim \mathcal{N}(f(x), \sigma^2), \hat{g}(x) \sim \mathcal{N}(g(x), \sigma^2 I_d)\]

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

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

\[P_0 = P(\{X_i, f_0(X_i), g_0(X_i)\}_{i=1}^{T}) \quad P_1 = P(\{X_i, f_1(X_i), g_1(X_i)\}_{i=1}^{T})\]
Lower bounds based on active learning

\[
\sup_{O} \sup_{S} \inf_{\hat{x}} \sup_{f} \mathbb{E}[\|\hat{x} - x_f^*\|] = \Omega(T^{-\frac{1}{2\kappa-2}})
\]

- **Fano’s Inequality**
  \[
  \inf_{\hat{x}} \sup_{f} P(\|\hat{x} - x_f^*\| > \|x_{f_0}^* - x_{f_1}^*\|/2) \geq \text{constant}
  \]

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

\[
= O(Ta^{2\kappa-2}) + O(Ta^{2\kappa})
\]

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

Query that yields max difference between function/gradient values

Castro-Nowak’07
Lower bounds based on active learning

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

- Fano’s Inequality
  \[ \inf \sup P(\|\hat{x} - x^*_f\| > \|x^*_0 - x^*_1\|/2) \geq \text{constant} \]

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

Query that yields max difference between function/gradient values

\[ = O(Ta^{2\kappa-2}) + O(Ta^{2\kappa}) \]

Also yields lower bounds for uniformly convex functions and zeroth-order oracle which match Iouditski-Nesterov’10, Jamieson-Nowak-Recht’12
Epoch-based gradient descent

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

\textbf{until} Oracle budget $T$ is exhausted \( \sum_{i=1}^e T_i \leq T \)

\textbf{for} $t = 1$ to $T_e$ do

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

$x_{1}^{e+1} = \frac{1}{T_e} \sum_{t=1}^{T_e} x_t^e$

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

\textbf{requires knowledge of} $\kappa$

\textbullet \ If $f$ is a convex function that satisfies TNC($\kappa$) and is Lipschitz

\( \sup_{O} \sup_{S} \inf_{\hat{x}} \sup_{f} \mathbb{E}[\|\hat{x} - x^*\|] = \tilde{O}(T^{-\frac{1}{2\kappa-2}}) \)

\( \sup_{O} \sup_{S} \inf_{\hat{x}} \sup_{f} \mathbb{E}[f(\hat{x}) - f(x^*)] = \tilde{O}(T^{-\frac{\kappa}{2\kappa-2}}) \)
Adapting to degree of convexity
Adapting to degree of convexity

For $e = 1, \ldots, E = \log \sqrt{T/\log T}$

Run any optimization procedure that is optimal for convex functions, with sample budget $T_e = T/E$

$R_{e+1} = R_e/2, e \leftarrow e + 1$

Adapted from Iouditski-Nesterov’10

\[ \exists \tilde{e} \text{ s.t. } ||x_{\tilde{e}} - x_{\tilde{e}}^*|| \leq T^{-1/(2\kappa-2)} \]

since

\[ \lambda ||x_e - x_e^*||^\kappa \leq f(x_e) - f(x_e^*) \leq \frac{R_e}{\sqrt{T}} \]

rate for convex Lipschitz functions

Also,

\[ x_{\tilde{e}}^* = x^* \]
Adapting to degree of convexity

For $e = 1, \ldots, E = \log \sqrt{T/\log T}$

Run any optimization procedure that is optimal for convex functions, with sample budget $T_e = T/E$

$R_{e+1} = R_e / 2, e \leftarrow e + 1$

Adapted from Iouditski-Nesterov’10

\[ \exists \bar{e} \text{ s.t. } \| x_{\bar{e}} - x_{\bar{e}}^* \| \leq T^{-1/(2\kappa-2)} \]

$x_{\bar{e}}^* = x^*$
Adapting to degree of convexity

For $e = 1, \ldots, E = \log \sqrt{T/\log T}$

Run any optimization procedure that is optimal for convex functions, with sample budget $T_e = T/E$

$R_{e+1} = R_e/2, e \leftarrow e + 1$

Adapted from Iouditski-Nesterov’10

$\exists \bar{e} \text{ s.t. } \|x_{\bar{e}} - x_{\bar{e}}^*\| \leq T^{-1/(2\kappa - 2)}$

$x_{\bar{e}}^* = x^*$
Adapting to degree of convexity

For $e = 1, \ldots, E = \log \sqrt{T / \log T}$

Run any optimization procedure that is optimal for convex functions, with sample budget $T_e = T/E$

$R_{e+1} = R_e / 2, e \leftarrow e + 1$

Adapted from Iouditski-Nesterov’10

\[ \exists \bar{e} \text{ s.t. } \|x_{\bar{e}} - x_{\bar{e}}^*\| \leq T^{-1/(2\kappa-2)} \]

\[ x_{\bar{e}}^* = x^* \]
Adapting to degree of convexity

For $e = 1, \ldots, E = \log \sqrt{T/\log T}$

Run any optimization procedure that is optimal for convex functions, with sample budget $T_e = T/E$

$R_{e+1} = R_e / 2, e \leftarrow e + 1$

Adapted from Iouditski-Nesterov’10

$\exists \bar{e} \ s.t. \ ||x_{\bar{e}} - x^*|| \leq T^{-1/(2\kappa-2)}$

$x^*_{\bar{e}} = x^*$

$\forall e \geq \bar{e}, \ ||x_e - x_{\bar{e}}|| \leq T^{-1/(2\kappa-2)}$
Adapting to degree of convexity

For $e = 1, \ldots, E = \log \sqrt{T / \log T}$

Run any optimization procedure that is optimal for convex functions, with sample budget $T_e = T / E$

$R_{e+1} = R_e / 2, e \leftarrow e + 1$

- If $f$ is a convex function that satisfies TNC($\kappa$) and is Lipschitz

$$\sup_{O} \sup_{S} \inf_{\hat{x}} \sup_{f} \mathbb{E}[\|\hat{x} - x^*_f\|] = \tilde{O}(T^{-\frac{1}{2\kappa-2}})$$

$$\sup_{O} \sup_{S} \inf_{\hat{x}} \sup_{f} \mathbb{E}[f(\hat{x}) - f(x^*)] = \tilde{O}(T^{-\frac{\kappa}{2\kappa-2}})$$
Adaptive active learning
Adaptive 1-dim active learning

Robust Binary Search adaptive to $\kappa$

For $e = 1, \ldots, E = \log \sqrt{T/\log T}$

(ignoring $\kappa$)

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
Adaptive 1-dim active learning

Robust Binary Search adaptive to $\kappa$

For $e = 1, \ldots, E = \log \sqrt{T/\log T}$

(ignoring $\kappa$)

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

$\exists \bar{e} \text{ s.t. } \|x_{\bar{e}} - x_{\bar{e}}^*\| \leq T^{-1/(2\kappa-2)}$

since

$c\|x_e - x_e^*\|^\kappa \leq \mathrm{Risk}(x_e) - \mathrm{Risk}(x_e^*) \leq \frac{R_e}{\sqrt{T}}$ passive rate for threshold classifiers

Also,

$x_{\bar{e}}^* = x^*$
Adaptive 1-dim active learning

Robust Binary Search adaptive to $\kappa$

For $e = 1, \ldots, E = \log \sqrt{T/\log T}$

Do passive learning with sample budget $Te = T/E$

$Re_{e+1} = Re/2, e \leftarrow e + 1$

Adapted from Iouditski-Nesterov’10

$\exists \bar{e}$ s.t. $\|x_{\bar{e}} - x^{*}\| \leq T^{-1/(2\kappa-2)}$

$x^{*}_{\bar{e}} = x^{*}$

$\forall e \geq \bar{e}, \|x_{e} - x_{\bar{e}}\| \leq T^{-1/(2\kappa-2)}$

Much simpler than Hanneke’09
Reference & Future directions


- Reduction from d-dim convex optimization to 1-dim active learning for $\kappa$-TNC functions ($\kappa \neq 2$)?
- Adaptive d-dimensional active learning/Model selection in active learning?
- Porting active learning results to yield non-convex optimization guarantees?

http://www.cs.cmu.edu/~aarti/