Cambridge Statistics Seminar
November 4, 2016

Constrained and Localized Nonparametric Estimation and Optimization

John Lafferty
Department of Statistics &
Department of Computer Science
University of Chicago
Sabyasachi Chatterjee
Chicago

John Duchi
Stanford

Yuancheng Zhu
Chicago / Wharton
Problem 1: Quantized Estimation

- We’re on different planets. I estimate a model from data.
Problem 1: Quantized Estimation

- We’re on different planets. I estimate a model from data.
- I want to share my model with you, but have limited communication bandwidth.
Problem 1: Quantized Estimation

- We’re on different planets. I estimate a model from data.
- I want to share my model with you, but have limited communication bandwidth.
- I send you a simplified version of my model.
Problem 1: Quantized Estimation

- We’re on different planets. I estimate a model from data.
- I want to share my model with you, but have limited communication bandwidth.
- I send you a simplified version of my model.
- How does the accuracy of the received model degrade as the number of bits used in the communication decreases?
Problem 2: Per-Instance Complexity

• How hard is it to optimize \textit{this} function?
Problem 2: Per-Instance Complexity

- How hard is it to optimize this function?
Problem 2: Per-Instance Complexity

- How hard is it to optimize *this* function?
Problem 1: Quantized Estimation

- Kepler sends a compressed model estimate
- How does the accuracy of the received model degrade as the number of bits used in the communication decreases?

Figure 2. Kepler-29 lightcurves. *Upper panel:* the quarter-normalized, calibrated *Kepler* photometry (PA); *lower panel:* the detrended, normalized flux. The transit times of each planet are indicated by dots at the bottom of each panel.

Fabrycky et al., 2012.
Climate simulations
Broader context

- We’re trying to understand statistical estimation under various constraints—shape, computational cost, storage cost.
- What can we say about tradeoffs between accuracy and the constraints?
Suppose we represent the estimator with $B$ bits. In an optimal representation, how much do we lose in terms of minimax risk?

We answer this by revisiting nonparametric minimax theory from perspective of rate-distortion theory.
Normal means model

Normal means is the archetypal nonparametric problem. Captures essentials of nonparametric estimation.

Observe $X_i \sim N(\theta_i, \sigma_i^2)$, for $i = 1, 2, \ldots, n$.

Goal: Estimate means $\theta_i$ to minimize the risk $R(\hat{\theta}, \theta) = \mathbb{E}\|\hat{\theta} - \theta\|^2$. 
Quantized normal means

We wish to limit the number of bits in our estimator:

\[ X_1, X_2, \ldots, X_n \mapsto \tilde{\theta}_1, \tilde{\theta}_2, \ldots, \tilde{\theta}_n \]

Classical rate-distortion setting:

minimize number of bits subject to \( E(X - \bar{X})^2 \leq D \)

In our estimation setting:

minimize number of bits subject to \( \inf_{\tilde{\theta}} E(\tilde{\theta}(X) - \theta)^2 \leq R \)

We are quantizing with respect to the risk, or estimation error — the distortion in our estimation of an unknown constant
General lower bound strategy

Prior $\pi(\theta)$ supported on $\Theta$, posterior mean $\delta(X_{1:n})$.

Integrated risk decomposes:

$$\int_\Theta E_\theta \|\theta - \dot{\theta}_n\|^2 d\pi(\theta) = E\|\theta - \delta(X_{1:n})\|^2 + E\|\delta(X_{1:n}) - \dot{\theta}_n\|^2$$
General lower bound strategy

Prior $\pi(\theta)$ supported on $\Theta$, posterior mean $\delta(X_{1:n})$.

Integrated risk decomposes:

$$\int_\Theta \mathbb{E}_\theta \|\theta - \hat{\theta}_n\|^2 d\pi(\theta) = \mathbb{E}\|\theta - \delta(X_{1:n})\|^2 + \mathbb{E}\|\delta(X_{1:n}) - \hat{\theta}_n\|^2$$

↑
Bayes risk
General lower bound strategy

Prior $\pi(\theta)$ supported on $\Theta$, posterior mean $\delta(X_{1:n})$.

Integrated risk decomposes:

$$\int_{\Theta} \mathbb{E}_\theta \|\theta - \tilde{\theta}_n\|^2 d\pi(\theta) = \mathbb{E}\|\theta - \delta(X_{1:n})\|^2 + \mathbb{E}\|\delta(X_{1:n}) - \tilde{\theta}_n\|^2$$

↑ Bayes risk ↑ excess risk from quantization
General lower bound strategy

With $T_n = T(X_1, \ldots, X_n)$ a sufficient statistic,

$$B_n \geq H(\hat{\theta}_n) \geq H(\hat{\theta}_n) - H(\hat{\theta}_n \mid \delta(T_n)) = I(\hat{\theta}_n; \delta(T_n))$$
General lower bound strategy

With \( T_n = T(X_1, \ldots, X_n) \) a sufficient statistic,

\[
B_n \geq H(\hat{\theta}_n) \geq H(\hat{\theta}_n) - H(\hat{\theta}_n \mid \delta(T_n)) = I(\hat{\theta}_n; \delta(T_n))
\]

Key optimization:

\[
\inf_{P(\cdot \mid \delta(T_n))} \mathbb{E}\|\delta(T_n) - \tilde{\theta}_n\|^2
\]

such that \( I(\tilde{\theta}_n; \delta(T_n)) \leq B_n \)
General lower bound strategy

With $T_n = T(X_1, \ldots, X_n)$ a sufficient statistic,

$$B_n \geq H(\tilde{\theta}_n) \geq H(\tilde{\theta}_n) - H(\tilde{\theta}_n \mid \delta(T_n)) = I(\tilde{\theta}_n; \delta(T_n))$$

Key optimization:

$$\inf_{P(\cdot \mid \delta(T_n))} \mathbb{E} \| \delta(T_n) - \tilde{\theta}_n \|^2$$

such that $I(\tilde{\theta}_n; \delta(T_n)) \leq B_n$

Denote value by $Q_n(\Theta, B_n; \pi)$. Then

$$R_n(\Theta, B_n) \geq \sup_{\pi} \left\{ R_n(\Theta; \pi) + Q_n(\Theta, B_n; \pi) \right\}$$
Shannon meets Kolmogorov

Donoho’s 1997 Wald Lectures show interplay between rate distortion, Kolmogorov’s metric entropy, and minimax theory.

<table>
<thead>
<tr>
<th>Shannon</th>
<th>Kolmogorov</th>
</tr>
</thead>
<tbody>
<tr>
<td>library</td>
<td>$X$ stochastic</td>
</tr>
<tr>
<td>representers</td>
<td>codebook $\mathcal{C}$</td>
</tr>
<tr>
<td>fidelity</td>
<td>$\mathbb{E} \min_{X' \in \mathcal{C}} | X - X' |^2$</td>
</tr>
<tr>
<td>complexity</td>
<td>$\log</td>
</tr>
</tbody>
</table>

$$H_\epsilon(\mathcal{F}) = \sup \left\{ R(\epsilon^2, X) : \mathbb{P}(X \in \mathcal{F}) = 1 \right\} (1 + o(1))$$
White noise model

Classical model for nonparametric regression:

\[ dY(t) = f(t)dt + \varepsilon dW(t) \]

where \( f \) lies in (periodic) Sobolev space

\[
\mathcal{W}_m(c) = \left\{ f \in L_2[0, 1] : \{\theta_j\} \in \Theta(m, c) \right\}
\]

where \( \theta_j = \langle f, \varphi_j \rangle \) for trigonometric basis, and \( \Theta(m, c) \) is ellipsoid

\[
\Theta(m, c) = \left\{ \theta : \sum_{j=1}^{\infty} j^{2m}\theta_j^2 \leq \frac{c^2}{\pi^{2m}} \right\}
\]
White noise model

Classical model for nonparametric regression:

\[ dY(t) = f(t)dt + \varepsilon dW(t) \]

where \( f \) lies in (periodic) Sobolev space

\[ \tilde{W}_m(c) = \{ f \in L_2[0, 1] : \{\theta_j\} \in \Theta(m, c) \} \]

where \( \theta_j = \langle f, \varphi_j \rangle \) for trigonometric basis, and \( \Theta(m, c) \) is ellipsoid

\[ \Theta(m, c) = \left\{ \theta : \sum_{j=1}^{\infty} j^{2m} \theta_j^2 \leq \frac{c^2}{\pi^{2m}} \right\} \]

We observe data

\[ Y_j = \int_0^1 \varphi_j(t)dY(t) = \theta_j + \varepsilon \xi_j \]

where \( \xi_j \sim N(0, 1) \).
Minimax risk for Sobolev ellipsoids

Minimax risk at noise level $\varepsilon$:

$$R_\varepsilon(m, c) = \inf \sup \mathbb{E}\|\hat{\theta} - \theta\|^2$$

Pinsker minimax bound:

$$R(m, c) = \lim \inf_{\varepsilon \to 0} \varepsilon^{-\frac{4m}{2m+1}} R_\varepsilon(m, c) \geq \left(\frac{c}{\pi m}\right)^\frac{2}{2m+1} (2m+1)^\frac{1}{2m+1} \left(\frac{m}{m+1}\right)^\frac{2m}{2m+1}$$
Minimax risk for Sobolev ellipsoids

Minimax risk at noise level $\varepsilon$:

$$R_{\varepsilon}(m, c) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta(m, c)} \mathbb{E}\|\hat{\theta} - \theta\|^2$$

Pinsker minimax bound:

$$R(m, c) = \liminf_{\varepsilon \to 0} \varepsilon^{\frac{-4m}{2m+1}} R_{\varepsilon}(m, c) \geq \left( \frac{c}{\pi m} \right)^{\frac{2}{2m+1}} (2m+1)^{\frac{1}{2m+1}} \left( \frac{m}{m+1} \right)^{\frac{2m}{2m+1}}$$

Pinsker constant $P_m(c)$
Quantized minimax risk for Sobolev ellipsoids

Minimax risk at noise level $\varepsilon$ and quantization level $B_\varepsilon$:

$$R_\varepsilon(m, c, B) = \inf_{\tilde{\theta} \in \mathcal{M}(B)} \sup_{\theta \in \Theta(m, c)} \mathbb{E} \| \tilde{\theta} - \theta \|^2$$
Quantized minimax risk for Sobolev ellipsoids

Minimax risk at noise level $\varepsilon$ and quantization level $B_\varepsilon$:

$$R_\varepsilon(m, c, B) = \inf_{\tilde{\theta} \in \mathcal{M}(B)} \sup_{\theta \in \Theta(m, c)} \mathbb{E}\|\tilde{\theta} - \theta\|^2$$

↑

estimators using $B$ bits
Quantized minimax risk for Sobolev ellipsoids

Minimax risk at noise level $\varepsilon$ and quantization level $B_\varepsilon$:

$$R_\varepsilon(m, c, B) = \inf_{\tilde{\theta} \in M(B)} \sup_{\theta \in \Theta(m, c)} \mathbb{E} \| \tilde{\theta} - \theta \|^2$$

↑

estimators using $B$ bits

Quantized minimax bound:

$$R(m, c, B) = \liminf_{\varepsilon \to 0} r(B_\varepsilon) R_\varepsilon(m, c, B_\varepsilon) \geq Q_m(c, B)$$
Quantized minimax risk for Sobolev ellipsoids

Minimax risk at noise level $\varepsilon$ and quantization level $B_\varepsilon$:

$$R_\varepsilon(m, c, B) = \inf_{\hat{\theta} \in \mathcal{M}(B)} \sup_{\theta \in \Theta(m, c)} \mathbb{E} \|\hat{\theta} - \theta\|^2$$

↑

estimators using $B$ bits

Quantized minimax bound:

$$R(m, c, B) = \liminf_{\varepsilon \to 0} r(B_\varepsilon) R_\varepsilon(m, c, B_\varepsilon) \geq Q_m(c, B)$$

↑

rate of convergence
Quantized minimax risk for Sobolev ellipsoids

Minimax risk at noise level $\varepsilon$ and quantization level $B_{\varepsilon}$:

$$R_{\varepsilon}(m, c, B) = \inf_{\tilde{\theta} \in \mathcal{M}(B)} \sup_{\theta \in \Theta(m,c)} \mathbb{E} \| \tilde{\theta} - \theta \|^2$$

↑

estimators using $B$ bits

Quantized minimax bound:

$$R(m, c, B) = \liminf_{\varepsilon \to 0} r(B_{\varepsilon}) R_{\varepsilon}(m, c, B_{\varepsilon}) \geq Q_m(c, B)$$

↑

rate of convergence

↑

Pinsker-Zhu constant
Regime change

Three regimes of quantization:

1. Lots of bits
2. Too few, suffer rate loss
3. Just enough to preserve rate (Goldilocks regime)
Quantized minimax estimation

Theorem.

1. If $B \varepsilon^{\frac{2}{2m+1}} \to \infty,$

$$\liminf_{\varepsilon \to 0} \varepsilon^{-\frac{4m}{2m+1}} R_\varepsilon(m, c, B) \geq P_m(c)$$
Quantized minimax estimation

**Theorem.**

1. If $B\varepsilon^{\frac{2}{2m+1}} \to \infty$,

   $$\liminf_{\varepsilon \to 0} \varepsilon^{-\frac{4m}{2m+1}} R_\varepsilon(m, c, B) \geq P_m(c)$$

2. If $B\varepsilon^{\frac{2}{2m+1}} \to 0$ and $B \to \infty$

   $$\liminf_{\varepsilon \to 0} B^{2m} R_\varepsilon(m, c, B) \geq \frac{c^2}{\pi^{2m}} m^{2m}$$
Theorem (continued)

3. If $B \varepsilon^{\frac{2}{2m+1}} \rightarrow d$,

$$\liminf_{\varepsilon \rightarrow 0} \varepsilon^{-\frac{4m}{2m+1}} R_\varepsilon(m, c, B) \geq Q_m(c, d)$$

where $Q_m(c, d)$ is the solution of a variational problem.
Recall: Lower bound strategy

Gaussian prior distribution $\theta_j \sim \mathcal{N}(0, \sigma_j^2)$. Bayes risk

$$\mathbb{E}\|\theta - \delta\|_2^2 = \sum_{j=1}^{\infty} \frac{\sigma_j^2 \varepsilon^2}{\sigma_j^2 + \varepsilon^2}.$$ 

Following our strategy, consider optimization

$$\inf \mathbb{E}\|\delta - \tilde{\theta}\|_2^2$$

such that $I(\delta; \tilde{\theta}) \leq B_{\varepsilon}$. 
Water filling

Define $\mu_j^2 = \mathbb{E}(\delta_j - \tilde{\theta}_j)^2$.

Classical “reverse water filling” argument gives

$$I(\delta; \tilde{\theta}) = \sum_{j=1}^{\infty} \frac{1}{2} \log_+ \left( \frac{\sigma_j^4}{\mu_j^2 (\sigma_j^2 + \varepsilon^2)} \right)$$

where $\log_+(x) = \max(\log x, 0)$. 
Lower bound variational problem

Thus,

\[ R_\varepsilon(m, c, B_\varepsilon) = \inf_{\hat{\theta}_\varepsilon, C(\hat{\theta}_\varepsilon) \leq B_\varepsilon} \sup_{\theta \in \Theta(m,c)} \mathbb{E}\|\theta - \hat{\theta}_\varepsilon\|^2 \geq Q_\varepsilon(B_\varepsilon, m, c)(1 + o(1)) \]

where \( Q_\varepsilon(B_\varepsilon, m, c) \) is value

\[
\max_{\sigma^2} \min_{\mu^2} \sum_{j=1}^{\infty} \frac{\sigma_j^2 \varepsilon^2}{\sigma_j^2 + \varepsilon^2} + \sum_{j=1}^{\infty} \mu_j^2
\]

such that

\[
\frac{1}{2} \sum_{j=1}^{\infty} \log_+ \left( \frac{\sigma_j^4}{\mu_j^2 (\sigma_j^2 + \varepsilon^2)} \right) \leq B_\varepsilon
\]

\[
\sum_{j=1}^{\infty} j^{2m} \sigma_j^2 \leq \frac{c^2}{\pi^{2m}}
\]
Sufficient regime: Excess risk due to quantization

![Graph showing the relationship between leading constant value and bits per coefficient d. The graph demonstrates the decrease in excess risk as the number of bits increases. The labels P+Q and Pinsker are used to indicate different curves on the graph.](image-url)
Achieving the lower bound: Intuition

- Generate “base code” shared between sender (Mars) and receiver (Earth)
Achieving the lower bound: Intuition

- Generate “base code” shared between sender (Mars) and receiver (Earth)
- Break up the base code in a way that is adaptive to the data. More bits used for blocks with larger signal size
Achieving the lower bound: Intuition

- Generate “base code” shared between sender (Mars) and receiver (Earth)
- Break up the base code in a way that is adaptive to the data. More bits used for blocks with larger signal size
- Normalize each block of the base code to get random vectors on unit sphere $\mathbb{S}^{T_k-1}$.

Find codeword $\tilde{Z}(k)$ having smallest angle to $Y(k)$. Transmit indices to Earth.

Reconstruct codewords on Earth, and form a type of quantized James-Stein estimator in each block.
Achieving the lower bound: Intuition

- Generate “base code” shared between sender (Mars) and receiver (Earth)
- Break up the base code in a way that is adaptive to the data. More bits used for blocks with larger signal size
- Normalize each block of the base code to get random vectors on unit sphere \( S^{T_k-1} \).
- Find codeword \( \tilde{Z}_{(k)} \) having smallest angle to \( Y_{(k)} \). Transmit indices to Earth.
Achieving the lower bound: Intuition

- Generate “base code” shared between sender (Mars) and receiver (Earth)
- Break up the base code in a way that is adaptive to the data. More bits used for blocks with larger signal size
- Normalize each block of the base code to get random vectors on unit sphere $\mathbb{S}^{T_k-1}$.
- Find codeword $\tilde{Z}_{(k)}$ having smallest angle to $Y_{(k)}$. Transmit indices to Earth.
- Reconstruct codewords on Earth, and form a type of quantized James-Stein estimator in each block.
Coding scheme

\[ Z_{ij} \sim \mathcal{N}(0, 1) \]
Coding scheme

\[ 2[T_1 \tilde{b}_1] \]

\[ 2[T_2 \tilde{b}_1] \]

\[ \ldots \]

\[ \ldots \]

\[ 2[T_K \tilde{b}_K] \]

\[ 2^B \]

\[ T_K \]
Coding scheme

\[ 2^{[T_1 b_1]} \]
\[ \tilde{Z}_1 \]

\[ 2^{[T_2 b_1]} \]
\[ \tilde{Z}_2 \]

\[ \ldots \]

\[ 2^{[T_K b_K]} \]
\[ \tilde{Z}_K \]

\( T_K \)

\( 2^B \)
Problem 1: Summary

- Computation-risk tradeoffs for communication/storage constraints can be sharply characterized
- Random coding schemes are exponential time. Future work: Coding/compression using sparse regression
- Many interesting directions possible: Other $L_p$ spaces, Besov spaces, quantized testing, distributed estimation, information complexity, ...

arXiv:1503.07368
Problem 2: Per-Instance Complexity

- How hard is it to estimate \textit{this} function?
- What is the complexity of optimizing \textit{this} function?
Heard around the Chicago Statistics lunch table

“Computer scientists are pessimists”

“I don’t care about minimax”
Heard around the Chicago Statistics lunch table

“Computer scientists are pessimists”

“I don’t care about minimax”

The pessimism is grounded in worst-case thinking, which is overly conservative.

(Theoretical) Statisticians are pessimists also. What are alternatives?
Minimax complexity of convex optimization

- \( \mathcal{F} \) class of convex functions on a convex set \( \mathcal{X} \subset \mathbb{R}^d \).

- \( \mathcal{O} \) a stochastic first-order oracle: query \((f, x) \in \mathcal{F} \times \mathcal{X}\), returns \( Z \in \mathbb{R}^d \), with mean \( f'(x) \in \partial f(x) \).

- \( \mathcal{A}_T \) class of all optimization methods that make \( T \) queries to \( \mathcal{O} \).

Nemirovski and Yudin (1983)
Minimax complexity of convex optimization

- $\mathcal{F}$ class of convex functions on a convex set $\mathcal{X} \subset \mathbb{R}^d$.

- $\mathcal{O}$ a stochastic first-order oracle: query $(f, x) \in \mathcal{F} \times \mathcal{X}$, returns $Z \in \mathbb{R}^d$, with mean $f'(x) \in \partial f(x)$.

- $\mathcal{A}_T$ class of all optimization methods that make $T$ queries to $\mathcal{O}$.

Minimax complexity

$$R_T(\mathcal{F}) = \inf_{A \in \mathcal{A}_T} \sup_{f \in \mathcal{F}} \mathbb{E}[\text{err}(f, A)] = \inf_{A \in \mathcal{A}_T} \sup_{f \in \mathcal{F}} \mathbb{E}\left[f(x_{T+1}) - \inf_{x \in \mathcal{X}} f(x)\right]$$

Nemirovski and Yudin (1983)
Minimax complexity for convex optimization

Known that

\[ R_T(\mathcal{F}_{sc}) \approx \frac{1}{T} \quad \text{strongly convex functions} \]

\[ R_T(\mathcal{F}_L) \approx \frac{1}{\sqrt{T}} \quad \text{Lipschitz functions} \]

Agarwal et al. (2010) extend analysis to \( d \)-dimensional case, also sparse setting
Raginsky and Rakhlin (2011) information theoretic proof technique; parallels minimax lower bounds in statistics.
Shortcomings of the framework?

- Ignores cost of computing the gradient – $O(1)$
- Does not allow for (decreasing) bias
- Does not account for computations on past gradients, e.g., quasi-Newton algorithms
- Too pessimistic and “global”
Per-Instance Complexity

What about a particular instance?

$\ll 1/T$?

$\approx 1/T$?

$\approx 1/\sqrt{T}$?
Local minimax complexity for optimization

Minimax complexity

\[
R_T(\mathcal{F}) = \inf_{A \in A_T} \sup_{f \in \mathcal{F}} \mathbb{E} \left[ \text{err}(f, A) \right] = \inf_{A \in A_T} \sup_{f \in \mathcal{F}} \mathbb{E} \left[ f(x_{T+1}) - \inf_{x \in \mathcal{X}} f(x) \right]
\]
Local minimax complexity for optimization

Minimax complexity

$$R_T(\mathcal{F}) = \inf_{A \in \mathcal{A}_T} \sup_{f \in \mathcal{F}} \mathbb{E}[\text{err}(f, A)] = \inf_{A \in \mathcal{A}_T} \sup_{f \in \mathcal{F}} \mathbb{E}\left[f(x_{T+1}) - \inf_{x \in \mathcal{X}} f(x)\right]$$

Local minimax complexity

$$R_T(f; \mathcal{F}) = \sup_{g \in \mathcal{F}} \inf_{A \in \mathcal{A}_T} \max\left\{ \mathbb{E}[\text{err}(f, A)], \mathbb{E}[\text{err}(g, A)] \right\}$$
Local minimax complexity

- Proposed by Cai and Low (2012, 2015) for shape-constrained estimation
- Quantifies difficulty of estimating this convex function $f$ at 0 (white noise model)

$$R_n(f) = \sup_{g \in \mathcal{F}} \inf_{\hat{T}_n} \max \left\{ R(\hat{T}_n, f), R(\hat{T}_n, g) \right\}$$

optimal risk for two model problem

- Hardest local alternative to $f$
Definitions 1/2

Set of minimum points: \( \mathcal{X}_f^* = \arg \min_{x \in C} f(x) \)

Error function: \( \text{err}(x, f) = \inf_{y \in \mathcal{X}_f^*} \|x - y\| \)

Minima separation: \( d(f, g) = \inf_{x \in \mathcal{X}_f^*, y \in \mathcal{X}_g^*} \|x - y\| \) for \( f, g \in \mathcal{F} \).

Exclusion inequality:

\[
\text{err}(x, f) < \frac{1}{2} d(f, g) \quad \Rightarrow \quad \text{err}(x, g) \geq \frac{1}{2} d(f, g).
\]
Definitions 2/2

\( f'(x) \in \partial f(x) \) unique subgradient returned as mean by the oracle when queried with \( x \). For example, \( f'(x) = \arg \min_{z \in \partial f(x)} \| z \| \). Define

\[
\kappa(f, g) = \sup_{x \in C} \| f'(x) - g'(x) \|
\]
Definitions 2/2

$f'(x) \in \partial f(x)$ unique subgradient returned as mean by the oracle when queried with $x$. For example, $f'(x) = \arg \min_{z \in \partial f(x)} \|z\|$. Define

$$\kappa(f, g) = \sup_{x \in C} \|f'(x) - g'(x)\|$$

Two dissimilarity measures between $f$ and $g$:

- $d(f, g)$ distance between minimizers
- $\kappa(f, g)$ largest separation between subgradients
Definitions 2/2

\( f'(x) \in \partial f(x) \) unique subgradient returned as mean by the oracle when queried with \( x \). For example, \( f'(x) = \arg \min_{z \in \partial f(x)} \| z \| \). Define

\[
\kappa(f, g) = \sup_{x \in C} \| f'(x) - g'(x) \| 
\]

Two dissimilarity measures between \( f \) and \( g \):

\[
d(f, g) \quad \text{distance between minimizers}
\]

\[
\kappa(f, g) \quad \text{largest separation between subgradients}
\]

Modulus of continuity of \( d \) with respect to \( \kappa \) at function \( f \):

\[
\omega_f(\varepsilon) = \sup_{g \in \mathcal{F}} \left\{ d(f, g) : \kappa(f, g) \leq \varepsilon \right\}
\]
Modulus of continuity

\[ \omega_f(\varepsilon) = \sup \left\{ \inf_{x \in \mathcal{X}_f^*} |x - y| : y \in \mathcal{C}, |f'(y)| < \varepsilon \right\} \]
Modulus of continuity

\[ \omega_f(\varepsilon) = \sup \left\{ \inf_{x \in X^*_f} |x - y| : y \in C, |f'(y)| < \varepsilon \right\} \]

\[ f(x) = \frac{1}{\alpha} |x|^\alpha \]

\[ \omega_f(\varepsilon) = \varepsilon^{\frac{1}{\alpha - 1}} \]

\[ f(x) = \begin{cases} 
\frac{1}{\alpha} |x|^\alpha & x \leq 0 \\
\frac{1}{\beta} |x|^\beta & x > 0 
\end{cases} \]

\[ \omega_f(\varepsilon) = \varepsilon^{\frac{1}{\alpha \vee \beta - 1}} \]
Modulus characterizes local minimax

**Theorem.** For all sufficiently large $T$,

$$C_1 \omega_f \left( \frac{\sigma}{\sqrt{T}} \right) \leq R_T(f; \mathcal{F}) \leq C_2 \omega_f \left( \frac{\sigma}{\sqrt{T}} \right).$$
Modulus characterizes local minimax

**Theorem.** For all sufficiently large $T$,

$$C_1 \omega_f \left( \frac{\sigma}{\sqrt{T}} \right) \leq R_T(f; \mathcal{F}) \leq C_2 \omega_f \left( \frac{\sigma}{\sqrt{T}} \right).$$

- For $f(x) = c|x - x^*|^{\alpha}$, error decays as $O(T^{-1/2(\alpha-1)})$.
- Proof works for any $d(f, g)$ satisfying exclusion inequality.
- Does not guarantee an algorithm that achieves this for any $f$. 
Superefficiency

Suppose an algorithm $A \in \mathcal{A}_T$ outperforms the modulus:

$$\mathbb{E}_f \text{err}(\hat{x}_A, f) \leq \delta_T \omega_f \left( \frac{\sigma}{\sqrt{T}} \right),$$

with $\delta_T \to 0$, $e^T \delta_T \to \infty$. Then exists functions with $\kappa(f, g_T) \to 0$ and

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{g_T} \text{err}(\hat{x}_A, g_T)}{\omega_{g_T} \left( \sigma \sqrt{T^{-1} \log(1/\delta_T)} \right)} > 0.$$
Superefficiency

Suppose an algorithm $A \in \mathcal{A}_T$ outperforms the modulus:

$$
\mathbb{E}_f \text{err}(\hat{x}_A, f) \leq \delta_T \omega_f \left( \frac{\sigma}{\sqrt{T}} \right),
$$

with $\delta_T \to 0$, $e^T \delta_T \to \infty$. Then exists functions with $\kappa(f, g_T) \to 0$ and

$$
\liminf_{T \to \infty} \frac{\mathbb{E}_{g_T} \text{err}(\hat{x}_A, g_T)}{\omega_{g_T} \left( \sigma \sqrt{T^{-1} \log(1/\delta_T)} \right)} > 0.
$$

Thus $\omega_f$ can be viewed as analogue of Fisher information for stochastic convex optimization.
Binary search algorithm

Input: $T, r$

Initialize: $(a_0, b_0), E = \lfloor r \log T \rfloor, T_0 = \lfloor T/E \rfloor,$

for $e = 1, \ldots, E$ do:

Query $x_e = (a_e + b_e)/2$ for $T_0$ times to get $Z_t^{(e)}$ for $t = 1, \ldots, T_0$

Calculate the average $\bar{Z}_{T_0}^{(e)} = \frac{1}{T_0} \sum_{t=1}^{T_0} Z_t^{(e)}$

If $\bar{Z}_{T_0}^{(e)} > 0$, set $(a_{e+1}, b_{e+1}) = (a_e, x_e)$

If $\bar{Z}_{T_0}^{(e)} \leq 0$, set $(a_{e+1}, b_{e+1}) = (x_e, b_e)$

end

Output: $x_E$
Binary search achieves the benchmark

**Theorem.** With probability at least $1 - \delta$ and for large enough $T$, 

$$\inf_{x \in X_f^*} |x_E - x| \leq \tilde{C} \omega_f \left( \frac{\sigma}{\sqrt{T}} \right)$$

where the term $\tilde{C}$ hides a dependence on $\log T$ and $\log(1/\delta)$. 
Binary search analysis

Proposition. For \( \delta \in (0, 1) \), let \( C_\delta = \sigma \sqrt{2 \log(E/\delta)} \). Define

\[
I_\delta = \left\{ y \in \text{dom}(f) : |f'(y)| < \frac{C_\delta}{\sqrt{T_0}} \right\}.
\]

Then with probability at least \( 1 - \delta \),

\[
\text{dist}(x_E, I_\delta) \leq 2^{-E}(b_0 - a_0).
\]
Simulations

Rates for $f(x) = |x - x^*|^k$ agree with adaptive estimation of uniformly convex functions (Juditsky and Nesterov, 2014)
Problem 2: Summary

- Framework for assessing complexity of minimizing individual convex functions.
- Close connections to classical statistical theory; not clear it’s “the correct” formulation.
- Outstanding problems: scaling and adaptive algorithm in higher dimensions

arXiv:1605.07596