Two Useful Arrows Darts in that Quiver

Clément Canonne

FOCS Workshop – November 9, 2019
Avering, Bucketing, and Investing arguments
Suppose you have $a: X \to [0, 1]$ such that

$$\mathbb{E}[a(x)] \geq \varepsilon.$$ 

(Let’s say you already proved that.) We think of $a(x)$ as the quality of $x$, and “using” it has cost $\text{cost}(a(x))$. 
Suppose you have $a: X \to [0, 1]$ such that

$$\mathbb{E}[a(x)] \geq \varepsilon.$$ 

(Let’s say you already proved that.) We think of $a(x)$ as the quality of $x$, and “using” it has cost $\text{cost}(a(x))$.

For instance, a population of coins, each with their own bias. The expected bias is $\varepsilon$; for any given coin, checking bias 0 vs. bias $\alpha$ takes $1/\alpha^2$ tosses. Goal: find a biased coin.
How... to convert this into a *useful* thing? How to find an $x$ with small cost?

That is, can we get

$$\Pr_x [ a(x) \geq \text{blah}(\varepsilon) ] \geq \text{bluh}(\varepsilon)$$

for some “good” functions $\text{blah}$, $\text{bluh}$?
“By a standard averaging argument…”

First attempt: Markov

Lemma (Markov)

We have

\[ \Pr \left[ a(x) \geq \frac{\varepsilon}{2} \right] \geq \frac{\varepsilon}{2}. \]  

(1)
“By a standard averaging argument...”

First attempt: Markov

Lemma (Markov)

We have

\[ \Pr_x \left[ a(x) \geq \frac{\epsilon}{2} \right] \geq \frac{\epsilon}{2}. \]  (1)

Proof.

\[ \epsilon \leq \mathbb{E}[a(x)] \leq \frac{\epsilon}{2} \cdot \Pr_x \left[ a(x) < \frac{\epsilon}{2} \right] + \underbrace{1 \cdot \Pr_x \left[ a(x) \geq \frac{\epsilon}{2} \right]}_{\leq 1} \]
“By a standard averaging argument…”

First attempt: Markov

**Strategy**

Sample $O(1/\varepsilon)$ $x$’s to find a “good” one; for each, pay cost($\varepsilon/2$).
“By a standard averaging argument…”

First attempt: Markov

**Strategy**
Sample $O(1/\varepsilon)$ $x$’s to find a “good” one; for each, pay cost($\varepsilon/2$).

**Yes, but…**
Typically, at least quadratic total cost in $\varepsilon$ as cost($\alpha$) = $\Omega(1/\alpha)$. 
“By a standard averaging argument…”
First attempt: Markov

Strategy
Sample $O(1/\varepsilon)$ $x$’s to find a “good” one; for each, pay cost($\varepsilon/2$).

Yes, but…
Typically, at least quadratic total cost in $\varepsilon$ as cost($\alpha$) = $\Omega(1/\alpha)$.
We should not pay the worst of both worlds.
“By a standard bucketing argument…”

Second attempt: my bucket list

**Lemma (Bucketing)**

There exists $1 \leq j \leq \lceil \log(2/\varepsilon) \rceil := L$ s.t.

$$
\Pr_x \left[ a(x) \geq 2^{-j} \right] \geq \frac{2^j \varepsilon}{4L}.
$$

(2)
“By a standard bucketing argument...”

Second attempt: my bucket list

Lemma (Bucketing)

There exists $1 \leq j \leq \lceil \log(2/\varepsilon) \rceil := L$ s.t.

$$\Pr_x \left[ a(x) \geq 2^{-j} \right] \geq \frac{2^j \varepsilon}{4L}.$$  \hfill (2)

Proof.

Define buckets $B_0 := \{x : a(x) \leq \varepsilon/2\}$,

$$B_j := \{x : 2^{-j} \leq a(x) \leq 2^{-j+1}\}, 1 \leq j \leq L$$

Then

$$\varepsilon \leq \mathbb{E}[a(x)] \leq \frac{\varepsilon}{2} \cdot \Pr \left[ x \in B_0 \right] + \sum_{j=1}^{L} 2^{-j+1} \cdot \Pr \left[ x \in B_j \right] \leq 1$$

so (averaging!) there exists $j^*$ s.t. $2^{-j+1} \cdot \Pr \left[ x \in B_j \right] \geq \varepsilon/(2L)$.
Strategy

For each $j \in [L]$, in case it’s the good bucket:

+ sample $O(\log(1/\epsilon)/(2^j \epsilon))$ x’s to find a “good” one in $B_j$;
+ for each such $x$, pay cost($2^{-j}$).

“By a standard bucketing argument…”
Second attempt: my bucket list
“By a standard bucketing argument…”

Second attempt: my bucket list

Strategy

For each $j \in [L]$, in case it’s the good bucket:

- sample $O(\log(1/\varepsilon)/(2^j \varepsilon))$ x’s to find a “good” one in $B_j$;
- for each such $x$, pay cost($2^{-j}$).

Total cost (examples):

$$\sum_{j=1}^{L} \frac{\log(1/\varepsilon)}{2^j \varepsilon} \cdot \text{cost}(2^{-j}) = \begin{cases} \frac{\log^2(1/\varepsilon)}{\varepsilon} & \text{if } \text{cost}(\alpha) \asymp 1/\alpha \\ \frac{\log(1/\varepsilon)}{\varepsilon^2} & \text{if } \text{cost}(\alpha) \asymp 1/\alpha^2 \end{cases}$$
“By a standard bucketing argument…”

Second attempt: my bucket list

**Strategy**

For each $j \in [L]$, **in case it’s the good bucket:**

- sample $O(\log(1/\varepsilon)/(2^j \varepsilon))$ $x$’s to find a “good” one in $B_j$;
- for each such $x$, pay cost($2^{-j}$).

**Total cost (examples):**

$$
\sum_{j=1}^{L} \frac{\log(1/\varepsilon)}{2^j \varepsilon} \cdot \text{cost}(2^{-j}) = \begin{cases} 
\frac{\log^2(1/\varepsilon)}{\varepsilon} & \text{if } \text{cost}(\alpha) \approx 1/\alpha \\
\frac{\log(1/\varepsilon)}{\varepsilon^2} & \text{if } \text{cost}(\alpha) \approx 1/\alpha^2
\end{cases}
$$

**Yes, but…**

we lose log factors. Do we *have* to lose log factors?
“By a refined averaging argument…”

Third (and last) attempt: strategic investment

Assume that \( \text{cost}(\alpha) \) is superlinear, e.g., \( \text{cost}(\alpha) = 1/\alpha^2 \).

Lemma (Levin’s Economical Work Investment Strategy)

There exists \( 1 \leq j \leq \lceil \log(2/\epsilon) \rceil := L \) s.t.

\[
\Pr_x \left[ a(x) \geq 2^{-j} \right] \geq \frac{2^j \epsilon}{8(L + 1 - j)^2}.
\]
“By a refined averaging argument…”

Third (and last) attempt: strategic investment

Assume that cost(\(\alpha\)) is superlinear, e.g., cost(\(\alpha\)) = \(1/\alpha^2\).

Lemma (Levin’s Economical Work Investment Strategy)

There exists \(1 \leq j \leq \lceil \log(2/\varepsilon) \rceil := L\) s.t.

\[
\Pr_x \left[ a(x) \geq 2^{-j} \right] \geq \frac{2^j \varepsilon}{8(L + 1 - j)^2}.
\]

Proof.

By contradiction:

\[
\mathbb{E}[a(x)] \leq \frac{\varepsilon}{2} + \sum_{j=1}^{L} 2^{-j+1} \cdot \Pr[ x \in B_j ] \leq \frac{\varepsilon}{2} + \sum_{j=1}^{L} 2^{-j+1} \cdot \Pr \left[ a(x) \geq 2^{-j} \right]
\]

\[
< \frac{\varepsilon}{2} + \sum_{j=1}^{L} 2^{-j+1} \cdot \frac{2^j \varepsilon}{8(L + 1 - j)^2} = \frac{\varepsilon}{2} + \frac{\varepsilon}{4} \sum_{\ell=1}^{L} \frac{1}{\ell^2} < \frac{\varepsilon}{2} + \frac{\varepsilon}{4} \sum_{\ell=1}^{\infty} \frac{1}{\ell^2} < \varepsilon
\]

“Oops.”
“By a refined averaging argument…”

Third (and last) attempt: strategic investment

**Strategy**

For each $j \in [L]$:

- sample $O((L + 1 - j)^2/(2^j \varepsilon))$ x’s to find a “good” one in $B_j$;
- for each such $x$, pay cost($2^{-j}$) $\simeq 2^{2j}$.
Strategy

For each $j \in [L]$:

- sample $O((L + 1 - j)^2 / (2^j \varepsilon))$ $x$’s to find a “good” one in $B_j$;
- for each such $x$, pay cost($2^{-j}$) $\approx 2^{2j}$.

Total cost:

$$\sum_{j=1}^{L} \frac{(L + 1 - j)^2}{2^j \varepsilon} \cdot 2^{2j} = \frac{1}{\varepsilon} \sum_{j=1}^{L} (L + 1 - j)^2 \cdot 2^j = \frac{2^{L+1}}{\varepsilon} \sum_{\ell=1}^{L} \ell^2 \cdot 2^{-\ell}$$

$$< \frac{4}{\varepsilon^2} \sum_{\ell=1}^{\infty} \ell^2 \cdot 2^{-\ell} \quad \text{(It’s 6.)}$$

O(1)
“By a refined averaging argument…”

Third (and last) attempt: strategic investment

**Strategy**

For each \( j \in [L] \):

- sample \( O((L + 1 - j)^2 / (2^j \epsilon)) \) x’s to find a “good” one in \( B_j \);
- for each such \( x \), pay cost \( (2^{-j}) \approx 2^j \).

**Total cost:**

\[
\sum_{j=1}^{L} \frac{(L + 1 - j)^2}{2^j \epsilon} \cdot 2^{2j} = \frac{1}{\epsilon} \sum_{j=1}^{L} (L + 1 - j)^2 \cdot 2^j = \frac{2^{L+1}}{\epsilon} \sum_{\ell=1}^{L} \ell^2 \cdot 2^{-\ell}
\]

\[
< \frac{4}{\epsilon^2} \sum_{\ell=1}^{\infty} \ell^2 \cdot 2^{-\ell} = O(1)
\]

(It’s 6.)

Yes, but…

No, actually, nothing. Works for any cost(\( \alpha \)) \( \gg 1 / \alpha^{1+\delta} \).
“By a refined averaging argument…”

Third (and last) attempt: strategic investment

Strategy

For each $j \in [L]$:

- sample $O((L + 1 - j)^2 / (2^j \varepsilon))$ x’s to find a “good” one in $B_j$;
- for each such $x$, pay cost$(2^{-j}) \asymp 2^j$.

Total cost:

$$\sum_{j=1}^{L} \frac{(L + 1 - j)^2}{2^j \varepsilon} \cdot 2^j = \frac{1}{\varepsilon} \sum_{j=1}^{L} (L + 1 - j)^2 \cdot 2^j \leq \frac{2^{L+1}}{\varepsilon} \sum_{\ell=1}^{L} \ell^2 \cdot 2^{-\ell}$$

$$< \frac{4}{\varepsilon^2} \sum_{\ell=1}^{\infty} \ell^2 \cdot 2^{-\ell} \quad \text{(It’s 6.)}$$

Yes, but…

No, actually, nothing. Works for any cost$(\alpha) \gg 1/\alpha^{1+\delta}$.

For cost$(\alpha) \asymp 1/\alpha$, not so easy, but some stuff exists.
Thomas’ Favorite Lemma
Kullback–Leibler Divergence

Recall the definition of Kullback–Leibler divergence (a.k.a. relative entropy) between two discrete distributions $p, q$:

$$D(p\|q) = \sum_\omega p(\omega) \log \frac{p(\omega)}{q(\omega)}$$
Kullback–Leibler Divergence

Recall the definition of Kullback–Leibler divergence (a.k.a. relative entropy) between two discrete distributions \( p, q \):

\[
D(p\|q) = \sum_{\omega} p(\omega) \log \frac{p(\omega)}{q(\omega)}
\]

It has some issues (symmetry, triangle inequality), yes, but it is everywhere (for a reason). It also has many nice properties.
Kullback–Leibler Divergence

The dual characterization

**Theorem (First)**

*For every* $q \ll p$,*

$$D(p\|q) = \sup_f \left( \mathbb{E}_{x\sim p} [f(x)] - \log \mathbb{E}_{x\sim q} [e^{f(x)}] \right)$$  \hspace{1cm} (4)

**Theorem (Second)**

*For every* $q \ll p$, *and every* $\lambda$

$$\log \mathbb{E}_{x\sim p} [e^{\lambda x}] = \max_{q \ll p} (\lambda \mathbb{E}_{x\sim q} [x] - D(q\|p))$$  \hspace{1cm} (5)

Known as: Gibbs variational principle (1902?), Donsker-Varadhan (1975), special case of Fenchel duality, …
Kullback–Leibler Divergence

The dual characterization

**Theorem (First)**

*For every* \( q \ll p \),

\[
D(p\|q) = \sup_{f} \left( \mathbb{E}_{x \sim p} [f(x)] - \log \mathbb{E}_{x \sim q} [e^{f(x)}] \right)
\]  \hspace{1cm} (4)

**Theorem (Second)**

*For every* \( q \ll p \), *and every* \( \lambda \)

\[
\log \mathbb{E}_{x \sim p} [e^{\lambda x}] = \max_{q \ll p} (\lambda \mathbb{E}_{x \sim q} [x] - D(q\|p))
\]  \hspace{1cm} (5)

Known as: Gibbs variational principle (1902?), Donsker-Varadhan (1975), special case of Fenchel duality, . . .
Theorem
Suppose $p$ is subgaussian on $\mathbb{R}^d$. For every function $a: \mathbb{R}^d \rightarrow [0, 1]$ (with $\alpha := \mathbb{E}_{x \sim p}[a(x)] > 0$),

$$\|\mathbb{E}_{x \sim p}[xa(x)]\|_2 \leq C_p \alpha \sqrt{\log \frac{1}{\alpha}}$$

(constant $C_p$ depends on subgaussian parameter, not on $d$).
An application

Theorem
Suppose $p$ is subgaussian on $\mathbb{R}^d$. For every function $a: \mathbb{R}^d \to [0, 1]$ (with $\alpha := \mathbb{E}_{x \sim p}[a(x)] > 0$),

$$\|\mathbb{E}_{x \sim p}[xa(x)]\|_2 \leq C_p \alpha \sqrt{\log \frac{1}{\alpha}} \hspace{1cm} (6)$$

(constant $C_p$ depends on subgaussian parameter, not on $d$).

The proof that follows was communicated to me by Himanshu Tyagi.
An application (and its proof, Gaussian case)

Setting \( z = x_i \) and \( q \ll p \) as \( \frac{d q}{d q}(x) = \frac{a(x)}{\mathbb{E}_p[a(x)]} \), we get

\[
\lambda \mathbb{E}_q[x_i] \leq \log \mathbb{E}_p \left[ e^{\lambda x_i} \right] + D(q_i \| p_i) = \frac{\lambda^2}{2} + D(q_i \| p_i),
\]

Optimizing for \( \lambda \), \( \mathbb{E}_q[x_i] \leq \sqrt{2D(q_i \| p_i)} \), i.e., \( \mathbb{E}_q[x_i]^2 \leq 2D(q_i \| p_i) \). Summing both sides over \( 1 \leq i \leq d \),

\[
\| \mathbb{E}_q[x] \|^2 \leq 2 \sum_{i=1}^{d} D(q_i \| p_i).
\]

and playing with nice properties of (conditional) relative entropy (chain rule, etc.) this is at most

\[
\sum_{i=1}^{d} \mathbb{E}_{x^{i-1}} \left[ D(q_{x_i|x^{i-1}} \| p_{x_i}) \right] = D(q \| p) = \frac{\mathbb{E}_p \left[ a(x) \log a(x) \right]}{\mathbb{E}_p[a(x)]} + \log \frac{1}{\mathbb{E}_p[a(x)]},
\]

which completes the proof. \( \square \)
I guess I’m done.