Adaptive Weighted Averaging

Machine Learning for Algorithms Workshop @STOC26

Aditya Bhaskara

(joint work with Ashok Cutkosky, Ravi Kumar, Manish Purohit)

THE UNIVERSITY of UTAH

Common problem in ML

Many models to choose from:

- pick the best combination for given task
- many sources of “advice”

Multiple models (different checkpoints, hyperparams, …)

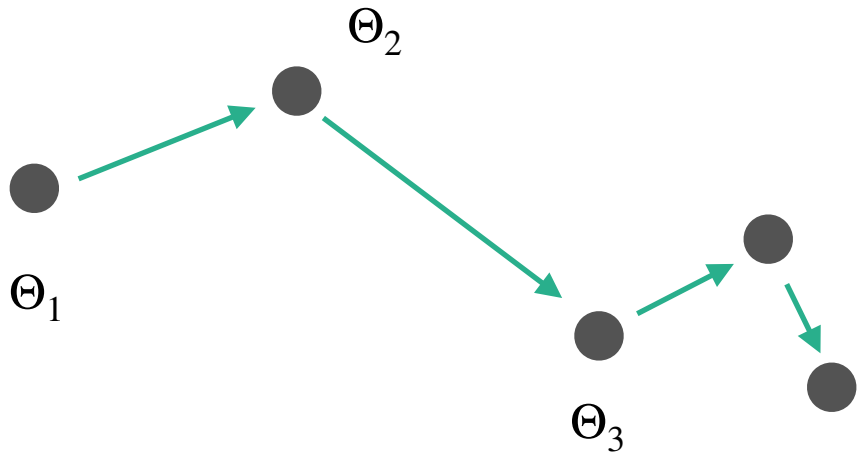
- how best to combine them?
- pick the best? uniform average?

Another common theme: iterate averaging

- Learning trajectory: $\Theta_1, \Theta_2, \ldots$
- Find one Θ that is "as good as the best"
- Standard approach: take the simple average

$$\frac{\Theta_1 + \Theta_2 + \ldots + \Theta_n}{n}$$

- Can we do better if the *best* is considerably better than average?



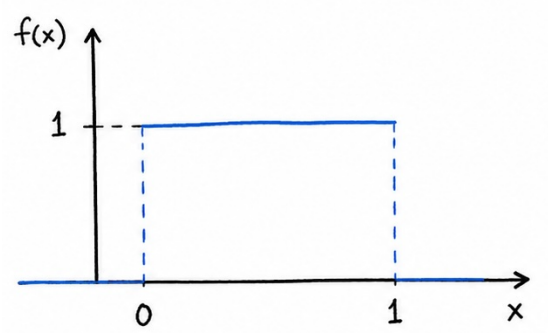
Too expensive to fully evaluate each model

Can a small sample help?

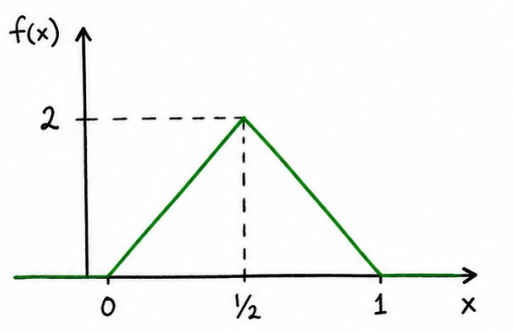
Distribution with the largest mean

Problem: *given a "few" samples from distributions* $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n$,

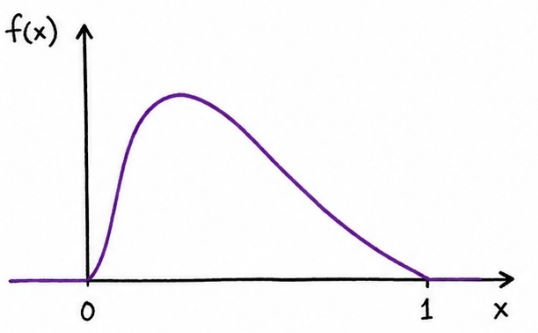
find the $\mathcal{D}_i$ *with the largest mean*

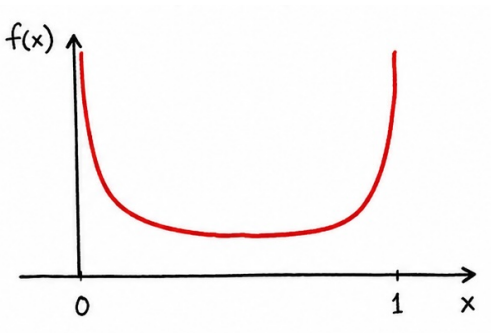


$\mathcal{D}_1$



$\mathcal{D}_2$





$\mathcal{D}_n$

values y_1^1, y_2^1, y_3^1

values y_1^2, y_2^2, y_3^2

…

Distribution with the largest mean

Problem: *given a "few" samples from distributions* $\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_n$,

find the $\mathscr{D}_i$ *with the largest mean*

- Will assume: Each $\mathscr{D}_i$ supported on [0,1], mean μ_i (unknown)
- Goal: output $z \in \Delta_n$ such that $\langle z, \mu \rangle = \sum_i z_i \mu_i$ is maximized

Isn't this just best arm identification? (full info)

One sample per distribution

Problem: ***given independent samples*** $y_1, y_2, \ldots, y_n$ ***from*** $\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_n$,

find $z \in \Delta_n$

- Adaptive strategy specified by a function $S : [0,1]^n \mapsto \Delta_n$
- Is there a natural notion of an “optimal” strategy?
- Can we find a strategy that *always* beats a (given) non-adaptive baseline?

Comparing strategies

Problem: ***given independent samples*** $y_1, y_2, \ldots, y_n$ ***from*** $\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_n$,

find $z \in \Delta_n$

- Adaptive strategy $S : [0,1]^n \mapsto \Delta_n$
- **Value** of a strategy: $\mathbb{E}_y \langle S(y), \mu \rangle$, where $y \sim \mathscr{D}_1 \times \mathscr{D}_2 \times \ldots \times \mathscr{D}_n$

Denoted **Val**(S)

Comparing strategies

- Given two strategies *S, T*, we say that S "dominates" T if for all distributions $\{\mathscr{D}_i\}_{i\in[n]}$ we have $\text{Val}(S) \geq \text{Val}(T)$

  ["Strictly" dominates if a strict inequality holds for at least one set of distributions.]

- Very strong requirement!

Fact: no strategy (possibly adaptive) can dominate the non-adaptive strategy that outputs (1,0,0,…,0)

Comparing strategies

- Given two strategies *S, T*, we say that S "dominates" T if for all distributions $\{\mathscr{D}_i\}_{i\in[n]}$ we have $\text{Val}(S) \geq \text{Val}(T)$

  ["Strictly" dominates if a strict inequality holds for at least one set of distributions.]

- Very strong requirement!

Fact: no strategy (possibly adaptive) can dominate the non-adaptive strategy that outputs (1,0,0,…,0)

Proof. Take any strategy S that dominates "e_1".
$\text{Val}(S) = \mathbb{E}_y\langle S(y), \mu\rangle = \langle \mathbb{E}_y S(y), \mu\rangle$.

Take $\mu = (0.9, \epsilon, \epsilon, \ldots, \epsilon)$ (adding to 1). If $S(y)$ is not $(1,0,0,\ldots,0)$ some y, we get a contradiction!

Comparing strategies

- Given two strategies *S*, *T*, we say that S "dominates" T if for all distributions $\{\mathscr{D}_i\}_{i\in[n]}$ we have $\text{Val}(S) \geq \text{Val}(T)$

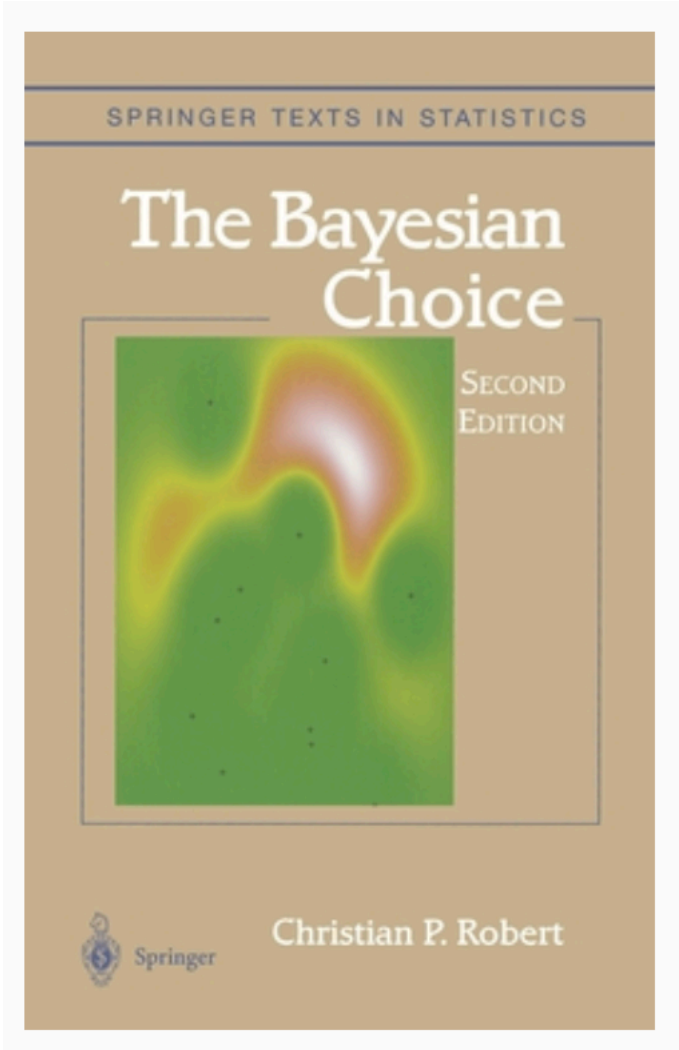
  ["Strictly" dominates if a strict inequality holds for at least one set of distributions.]

- Very strong requirement!

Fact: no strategy (possibly adaptive) can dominate the non-adaptive strategy that outputs (1,0,0,…,0)

Note: this is a very uninteresting strategy to dominate!

Well-studied subject



- A strategy S that is not strictly dominated by any other strategy T is called “admissible”
- We saw: the constant strategy $(1,0,\ldots,0)$ is admissible
- Interestingly, this is the *only* constant (or non-adaptive) strategy that is admissible

Non-adaptive strategies can be dominated

Theorem. *Let p be any "constant" strategy that has at least two non-zero entries. There exists an adaptive strategy that strictly dominates p.*

Quite powerful: same adaptive strategy helps (or doesn't hurt) for any set of (unknown) distributions $\mathscr{D}_1, \mathscr{D}_2, \ldots$

- Natural to ask: can we quantify the improvement in val($\cdot$) ?

- Simplest baseline: $p = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$ [uniform baseline]

Non-adaptive strategies can be dominated

Theorem. *Let p be any "constant" strategy that has at least two non-zero entries. There exists an adaptive strategy that strictly dominates p.*

Can we quantify the improvement (e.g., over uniform)?

- If μ_i are all equal, impossible to do better!
- "Variance" in the means is important
- When p is uniform, value improves by at least $\sigma^2(\mu)$ (defined shortly)

Candidate approach

Theorem. *Let p be any "constant" strategy that has at least two non-zero entries. There exists an adaptive strategy that strictly dominates p.*

The online learning approach:

- Start with distribution p
- Run one iteration of MWU

$p_1 \quad p_2 \quad \dots \quad p_n$

observe $y_1, y_2, \dots, y_n$

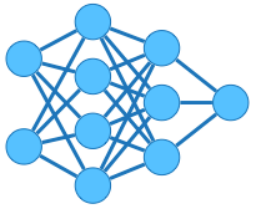
$p_1 \cdot e^{\eta y_1} \quad p_2 \cdot e^{\eta y_2} \quad \dots \quad (renormalize)$

Does not work! :(

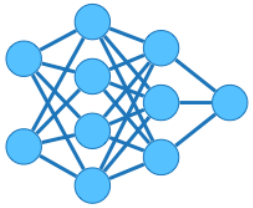
Simpler setting: p is uniform

Theorem. *Let* $p = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. *There exists an adaptive strategy that strictly dominates* p.

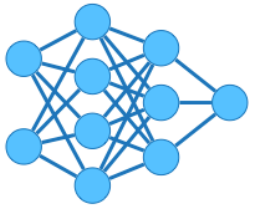
- Already non-trivial: val(S) $\geq$ val(Unif) for *all* $\{\mathscr{D}_i\}$
- For iterate averaging or model averaging — shows that using a single minibatch and getting loss estimates may be useful!



Θ_1



Θ_2



Θ_3

Simpler setting: p is uniform

Theorem. *Let* $p = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. *There exists an adaptive strategy that strictly dominates* p.

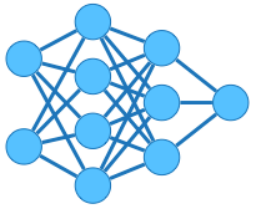
- Already non-trivial: val(S) $\geq$ val(Unif) for *all* $\{\mathscr{D}$
- For iterate averaging or model averaging — sho
  minibatch and getting loss estimates may

baseline:

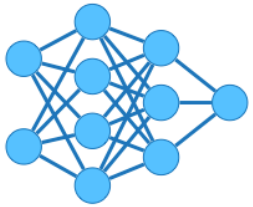
$$L\left(\frac{\Theta_1 + \Theta_2 + \Theta_3}{3}\right) \leq \frac{L(\Theta_1) + L(\Theta_2) + L(\Theta_3)}{3}$$

weighted:

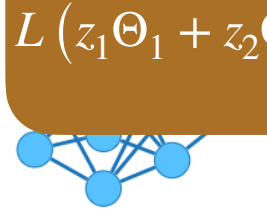
$$L\left(z_1\Theta_1 + z_2\Theta_2 + z_3\Theta_3\right) \leq z_1 L(\Theta_1) + z_2 L(\Theta_2) + z_3 L(\Theta_3)$$



Θ_1



Θ_2



Θ_3

Simpler setting: p is uniform

Theorem. *Let* $p = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. *There exists an adaptive strategy that strictly dominates* p.

- Seemingly toy— binary rewards: assume that $\mathscr{D}_i = \text{Bern}(x_i)$ (for all i)
- The samples $y = (y_1, y_2, \ldots, y_n) \in \{0,1\}^n$
- Every strategy S is determined by 2^n weight vectors $S((0,1,0,1)) = (0,\frac{1}{2},0,\frac{1}{2})$
- Given say $y = (0,1,0,1)$, what should S do?

Strategy: uniform over the 1's

Simpler setting: p is uniform

Theorem. *Let* $p = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ *and* $\mathcal{D}_i = Bern(x_i)$. *The strategy "uniform over ones" (denoted Ones) satisfies* $Val(Ones) \geq Val(p) + \frac{n}{n-1}\sigma^2(x)$, *where*

$$\sigma^2(x) := average\{(x_i - x_j)^2\}.$$

Proof relies on finding a polynomial expression for the "gain" and using a careful concavity argument

bound can be further improved when x_i are all small

Uniform p, arbitrary distributions

Theorem. *Let* $p = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$. *There exists an adaptive strategy that strictly dominates* p.

- What if $\mathcal{D}_i$ are arbitrary distributions? (still supported on [0,1], say)
- Now $S : [0,1]^n \mapsto \Delta_n$
- Given say $y = (0,\ 0.3,\ 0.7,\ 1)$, what should S do?
- "Overfitting" seems possible

Given $y = (0,\ 0.3,\ 0.7,\ 1)$,

what should S do?

Attempt inspired by rounding:

- post-hoc — convert y to a binary y', rounding each y_i to 1 with prob. y_i ("convert to Ber

- define $S(y)$ to be the *expecte*

directly implies that same "gain" statement holds for arbitrary $\mathscr{D}_i$!

Fun fact. *(Thanks Gemini) Obtained function is precisely the multi-linear extension of the "uniform on ones" function defined on* $\{0,1\}^n$

Uniform p, arbitrary distributions

Theorem. *Let* $p = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ *and* $\mathcal{D}_i$ *be* ***arbitrary****. The strategy "convert to Bernoulli" (denoted* S_{Bern}*) satisfies* $Val(S_{Bern}) \geq Val(Unif) + \frac{n}{n-1}\sigma^2(x)$*, where*

$$\sigma^2(x) := average\{(x_i - x_j)^2\}.$$

bound can be further improved when x_i are all small

Uniform p, arbitrary distributions

Theorem. *Let* $p = \left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$ *and* $\mathcal{D}_i$ *be* ***arbitrary****. The strategy "convert to Bernoulli" (denoted* S_{Bern}*) satisfies* $Val(S_{Bern}) \geq Val(Unif) + \frac{n}{n-1}\sigma^2(x)$*, where*

$$\sigma^2(x) := average\{(x_i - x_j)^2\}.$$

Theorem 2. *The convert-to-Bernoulli strategy is "admissible", i.e., no other strategy dominates it for all distributions.*

bound can be further improved when x_i are all small

Proof first shows that any dominating strategy must agree with S_{Bern} on $y \in \{0,1\}^n$; then uses careful induction to show impossibility of dominance

Arbitrary (non-adaptive) baselines

- Is there a strategy that dominates a given p ?
- Focus on the Bernoulli case: is there an analog of “uniform over ones”?

  Say we see values $y = (0, 1, 1, 1)$ and $p = (\frac{1}{9}, \frac{2}{3}, \frac{1}{9}, \frac{1}{9})$

Arbitrary baselines

- Is there a strategy that dominates any given p ?
- Focus on the Bernoulli case: is there an analog of "uniform over ones"?

  Say we see values $y = (0, 1, 1, 1)$ and $p = (\frac{1}{9}, \frac{2}{3}, \frac{1}{9}, \frac{1}{9})$

- Guess: $S(y) \propto (0, \frac{2}{3}, \frac{1}{9}, \frac{1}{9})$

Turns out this does not work!
What works? "peeling"

$$(\frac{1}{9}, \frac{2}{3}, \frac{1}{9}, \frac{1}{9}) = \frac{1}{9}(1,1,1,1) + \frac{5}{9}(0,1,0,0)$$

Extensions, open questions

- Can we simultaneously do better than two fixed baselines? (No!)
- Bounds in terms of other parameters of distributions?

  (variances, …)
- Distributions with unbounded supports (e.g., Gaussians)
- Right way to treat multiple samples?
  - we converge to the maximum μ, but rate?

Thank you!