Parameter-free online learning via model selection

Dylan Foster, Satyen Kale, Mehryar Mohri, and Karthik Sridharan

djfoster@cs.cornell.edu, satyenkale@google.com, mohri@cs.nyu.edu, sridharan@cs.cornell.edu

Cornell University Google NYU

Overview

Structural risk minimization for online convex optimization

or

How to do large-scale online/stochastic optimization without hyperparameters.

OCO and Online Gradient Descent

Online convex optimization protocol:
For $t = 1$ to n:

Select distribution $q_t \in \Delta(\mathcal{W})$ (where $\mathcal{W} \subseteq \mathbb{R}^d$ is constraint set).

Nature selects convex function $g_t : \mathcal{W} \to \mathbb{R}$.

Draw $w_t \sim q_t$ and incur loss $g_t(w_t)$.

End
Standard algorithm: online gradient descent. Suppose:

- $\mathcal{W} = \{w \in \mathbb{R}^d \mid \|w\|_2 \le R\}$ and each g_t is 1-Lipschitz wrt $\|\cdot\|_2$.
- Predict with Online Gradient Descent [1]:

$$w_{t+1} = \mathrm{Proj}_{\mathcal{W}}(w_t - \eta \nabla g_t(w_t)),$$

with $\eta = R/\sqrt{n}$.

OGD has regret:

$$\boxed{\sum_{t=1}^n g_t(w_t) - \inf_{w\in\mathcal{W}} \sum_{t=1}^n g_t(w) \le R\sqrt{n}.}$$

How to choose R?

Parameter-free learning

Solution [2, 3, 4, 5]: There are efficient (linear-time) algorithm achieving:

$$\sum_{t=1}^n g_t(w_t) - \sum_{t=1}^n g_t(w) \le (\|w\|_2 + 1)\sqrt{n \cdot \log((\|w\|_2+1)n)} \quad \forall w \in \mathbb{R}^d.$$

Same runtime as OGD + rate above is unimprovable.

This paper: Moving beyond ℓ_2 (efficiently)!

Results

Generalize to all norms

$$\sum_{t=1}^n g_t(w_t) - \sum_{t=1}^n g_t(w) \le (\|w\|+1)\sqrt{n \cdot \log((\|w\|+1)n)} \quad \forall w.$$

for any norm $\|\cdot\|$ where original (fixed-R) problem is learnable. Even ℓ_p analogue not known!

General structural bounds

$$\sum_{t=1}^n g_t(w_t) - \sum_{t=1}^n g_t(w) \le \mathbf{Comp}_n(w) \cdot \mathbf{Pen}(\mathbf{Comp}_n(w))$$

for abstract complexity $\mathbf{Comp}_n(w)$; allows arbitrary discrete or combinatorial structure.

Efficient meta-algorithm

– Efficient whenever original (parameter-dependent) problem has efficient algorithms.

$\Longrightarrow$ can work in non-convex or non-parametric settings.

Approach and key challenges

Theorem: Parameter-free mirror descent

Fix norm $\|\cdot\|$ with $\frac{1}{2}\|\cdot\|^2$ λ-strongly convex. Then parameter-free mirror descent efficiently guarantees

$$\mathbb{E}\left[\sum_{t=1}^n g_t(w_t) - \sum_{t=1}^n g_t(w)\right] \le (\|w\|+1)\sqrt{n \cdot \log((\|w\|+1)n)/\lambda} \quad \forall w.$$

whenever each g_t is 1-Lipschitz w.r.t. dual norm $\|\cdot\|_\star$.

Idea #1: Learn best learning rate for OMD

- Fix norm $\|\cdot\|$ with $\frac{1}{2}\|\cdot\|^2$ λ-strongly convex, let $\|\cdot\|_\star$ be the dual.
- Let $\mathcal{W}_k = \{w \in \mathbb{R}^d \mid \|w\| \le 2^{k-1}\}$, $k \in 1, \dots, n+1$.
- Then Online Mirror Descent over $\mathcal{W}_k$ guarantees

$$\sum_{t=1}^n g_t(w_t^k) - \inf_{w\in\mathcal{W}_k} \sum_{t=1}^n g_t(w) \le 2^{k-1}\sqrt{n/\lambda}$$

if $(g_t)_{t\le n}$ are 1-Lipschitz wrt $\|\cdot\|_\star$.

Idea #2: Reduce to experts problem

Recall experts setting over N experts:
For time $t = 1, \dots, n$:

- Learner selects distribution $p_t \in \Delta_N$.
- Nature selects loss $\boldsymbol{g}_t \in \mathbb{R}^N$.
- Learner samples $i_t \sim p_t$ and experiences loss $\boldsymbol{g}_t[i_t]$.

Regret:

$$\sum_{t=1}^n \mathbf{g}_t[i_t] - \min_{i\in[N]} \sum_{t=1}^n \mathbf{g}_t[i].$$

Applying to our setting:

- Experts: OMD instances $(w_t^k)_{k\in[N]}$ given above.
- Loss: $\mathbf{g}_t = (g_t(w_t^k))_{k\in[N]}$.
- Meta algorithm: Sample $i_t \sim p_t$ and play $w_t^{i_t}$.

Challenge:

- For our application, can have $|\mathbf{g}_t[i]| >> |\mathbf{g}_t[j]|$, e.g. 2^n vs. 2.
- Typical algorithms (eg: multiplicative weights) scale with $\|\mathbf{g}_t\|_\infty$.
- Can we ensure large coordinates don't dominate?

Idea #3: Multi-scale experts

Applying MultiscaleFTPL (see next column) to our OMD setting gives:

$$\mathbb{E}\left[\sum_{t=1}^n g_t(w_t^{k_t}) - \inf_{\|w\|\le 2^k} \sum_{t=1}^n g_t(w)\right] \le 2^k\sqrt{\frac{n}{\lambda}} + C\cdot 2^k\sqrt{n\log(2^k n)} \;\forall k \le n.$$

$\Longrightarrow$ **within constant factor of desired regret bound!**

- For $1 \le \|w\| \le 2^n$ the RHS is within a constant factor.
- Write off $\|w\| \le 1$.
- RHS of desired bound is vacuous for $\|w\| \ge 2^n$; no need to use algorithm.

Multi-scale experts algorithm: MultiscaleFTPL

Algorithm: MultiscaleFTPL

For time $t = 1, \dots, n$:

- Draw sign vectors $\sigma_{t+1}, \dots, \sigma_n \in \{\pm 1\}^N$ each uniformly at random.
- Compute distribution

$$p_t(\sigma_{t+1:n}) := \operatorname*{arg\,min}_{p\in\Delta_N} \sup_{\mathbf{g}_t : |\mathbf{g}_t[i]|\le c_i} \left[\langle p, \mathbf{g}_t\rangle + \sup_{i\in[N]}\left[-\sum_{s=1}^t \mathbf{g}_s[i] + 4\sum_{s=t+1}^n \sigma_s[i]c_i - B(i)\right]\right],$$

where $B(i) = 5c_i\sqrt{n\log(4c_i^2 n/\pi_i)}$

- Play $i_t \sim p_t$.

Theorem: Multi-scale experts

Suppose the loss sequence $(\mathbf{g}_t)_{t\le n}$ satisfies $|\mathbf{g}_t[i]| \le c_i$ for some $(c_i)_{i\in[N]}$ with each $c_i \ge 1$. Let $\pi \in \Delta_N$ be a given prior distribution on the experts. Then the randomized strategy MultiscaleFTPL has regret

$$\mathbb{E}\left[\sum_{t=1}^n \mathbf{g}_t[i_t] - \sum_{t=1}^n \mathbf{g}_t[i].\right] \le O\left(c_i\sqrt{n\log(nc_i/\pi_i)}\right) \quad \forall i \in [N].$$

!!! Compare to usual experts regret: $\|c\|_\infty\sqrt{n\log(1/\pi_i)}$. !!!

Recent applications of multi-scale experts:

- Online auctions [Bubeck-Devanur-Huang-Niazadeh'17]
  Uses alg. based on multi-scale entropy regularizer. Doesn't get optimal dependence on π_i but can give multiplicative regret guarantee.
- k-server [Bubeck-Cohen-Lee-Lee-Madry'17]
  Continuous-time generalization.

Analyzing MultiscaleFTPL

We will show that MultiscaleFTPL guarantees

$$\mathbb{E}\max_{i\in[N]}\left[\sum_{t=1}^n \mathbf{g}_t[i_t] - \sum_{t=1}^n \mathbf{g}_t[i] - B(i)\right] \le O(1).$$

Let's analyze worst case contribution of final round n to regret

$$\sup_{\mathbf{g}_n : |\mathbf{g}_n[i]|\le c_i} \mathbb{E}_{i_n\sim p_n} \max_{i\in[N]}\left[\mathbf{g}_n[i_n] - \sum_{t=1}^n \mathbf{g}_t[i] - B(i)\right].$$

Via (somewhat standard) technique of minimax swap + sequential symmetrization [Foster-Rakhlin-Sridharan'15], MultiscaleFTPL strategy leads to upper bound:

$$\sup_{\mathbf{g}_n : |\mathbf{g}_n[i]|\le c_i} \mathbb{E}_{\epsilon_n\in\{\pm1\}} \max_{i\in[N]}\left[2\epsilon_n \mathbf{g}_n[i_n] - \sum_{t=1}^{n-1} \mathbf{g}_t[i] - B(i)\right]$$

$$= \max_{\sigma_n\{\pm1\}^N} \mathbb{E}_{\epsilon_n\in\{\pm1\}} \max_{i\in[N]}\left[2\epsilon_n\sigma_n[i]c_i - \sum_{t=1}^{n-1} \mathbf{g}_t[i] - B(i)\right]$$

Multi-scale FTPL lemma says we can replace the $\max_{\sigma_n}$ with perturbation.

$$\le \mathbb{E}_{\sigma_n\{\pm1\}^N} \max_{i\in[N]}\left[4\sigma_n[i]c_i - \sum_{t=1}^{n-1} \mathbf{g}_t[i] - B(i)\right].$$

Repeating above process for n rounds (many subtleties — see paper), get upper bound on $\mathbf{Reg}_i - B(i)$ of

$$\mathbb{E}_{\sigma_{1:n}\{\pm1\}^N} \max_{i\in[N]}\left[4\sum_{t=1}^n \sigma_t[i]c_i - B(i)\right] \le O(1).$$

Final bound is by **multi-scale maximal inequality**.

Full result derived via adaptive relaxation framework [Foster-Rakhlin-Sridharan'15].

Key lemmas

Lemma: Multiscale Perturbation (cf. Rakhlin-Shamir-Sridharan'12))

For any $w \in \mathbb{R}^N$, any $c \in \mathbb{R}_+^N$,

$$\sup_{\sigma\in\{\pm1\}^N} \mathbb{E}_{\epsilon\in\{\pm1\}} \max_{i\in[N]}\{w_i + 2\epsilon\sigma_i c_i\} \le \mathbb{E}_{\sigma\in\{\pm1\}^N} \max_{i\in[N]}\{w_i + 4\sigma_i c_i\}.$$

Lemma: Multiscale Martingale Maximal Inequality

Let $(Z_t)_{t\le n}$ be any martingale difference sequence in $\mathbb{R}^N$ with $Z_t[i] \le c_i$ almost surely and $\pi \in \Delta_N$ be fixed. Then

$$\mathbb{E}_Z \sup_{i\in[N]}\left[2\sum_{t=1}^n Z_t[i] - 5c_i\sqrt{n\log(4c_i^2 n/\pi_i)}\right] \le O(1).$$

Supremum of scaled, offset random process.

- For each i, $|\sum_{t=1}^n Z_t[i]|$ is roughly $c_i\sqrt{n}$ whp.
- If we considered just $\mathbb{E}\sup_{i\in[N]}\sum_{t=1}^n Z_t[i]$, larger c_i terms would dominate.
- Offset $5c_i\sqrt{n\log(4c_i^2 n/\pi_i)}$ penalizes big c_is.

More applications

Online PCA task: Predict PSD matrix $W_t \in \mathbb{R}^{d\times d}$, receive PSD matrix Y_t with $\lambda_{\max}(Y_t) \le 1$, experience loss $\langle I - W_t, Y_t\rangle$.

Online PCA

There is a randomized algorithm for Online PCA that for all ranks $k \le d$ simultaneously achieves

$$\mathbb{E}\left[\sum_{t=1}^n \langle I - W_t, Y_t\rangle - \min_{\substack{W \text{ proj.} \\ \mathrm{rank}(W)=k}} \sum_{t=1}^n \langle I - W, Y_t\rangle\right] \le \widetilde{O}\left(\sqrt{n\min\{k, d-k\}^2}\right).$$

Suppose we're in same setting as parameter-free OMD, but want to adapt to multiple norms instead of a single one.

OCO with multiple norms

Fix a collection of N norms $\|\cdot\|_{(k)}$, each having $\frac{1}{2}\|\cdot\|_{(k)}^2$ λ_k-strongly convex. There is an efficient strategy that guarantees

$$\mathbb{E}\left[\sum_{t=1}^n g_t(w_t) - \sum_{t=1}^n g_t(w)\right] \le (\|w\|_{(k)}+1)\sqrt{n\cdot\log((\|w\|_{(k)}+1)n)/\lambda} \quad \forall w,\ k.$$

whenever each g_t is 1-Lipschitz w.r.t. each dual norm $\|\cdot\|_{(k),\star}$.

References

[1] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.

[2] Brendan McMahan and Jacob Abernethy. Minimax optimal algorithms for unconstrained linear optimization. In *Advances in Neural Information Processing Systems*, pages 2724–2732, 2013.

[3] H. Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In *Proceedings of The 27th Conference on Learning Theory*, pages 1020–1039, 2014.

[4] Francesco Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124, 2014.

[5] Francesco Orabona and Dávid Pál. From coin betting to parameter-free online learning. *arXiv preprint arXiv:1602.04128*, 2016.