Parameter-Free Convex Learning through Coin Betting

Francesco Orabona Dávid Pál
Yahoo Research, New York

YAHOO! RESEARCH

ARE YOU STILL TUNING HYPERPARAMETERS?

Regularized empirical risk minimization:

$$\arg\min_{w \in \mathbb{R}^d} \frac{\lambda}{2}\|w\|^2 + \sum_{i=1}^{N} f(w, x_i, y_i) \quad (1)$$

where f is convex in w.

- How do you choose the regularizer weight λ?

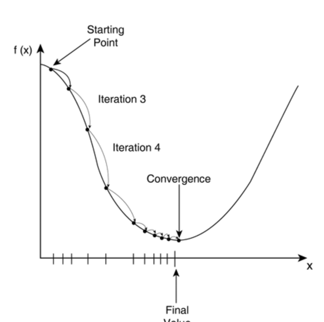
Stochastic approximation:

$$w_t = w_{t-1} - \eta_t \nabla f(w_{t-1}, x_t, y_t) \quad (2)$$

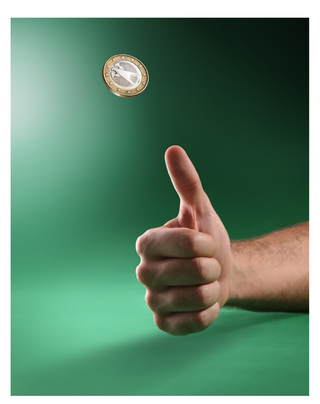
where f is convex in w.

- How do you choose the learning rate η_t?
- Why is the algorithm not able to select λ and/or η_t automatically?

FROM COIN-BETTING TO MACHINE LEARNING



is equivalent to



- Coin flip outcome $c_t \in \{+1, -1\}$.
- Krichevsky-Trofimov: Bet $\frac{1}{t}\sum_{i=1}^{t-1} c_i$ fraction of your current wealth on the most common outcome till time t.
- KT algorithm for coin betting gives rise to optimal parameter-free algorithms for Online Learning, Convex Optimization and Machine Learning!
- Key idea: Treat the gradient as the outcome of a coin flip.
- In other words: Learning rates are the results of suboptimal algorithms, they must be removed, not tuned/learned/adapted!

7 YEARS OF PARAMETER-FREE ALGORITHMS

- Streeter&McMahan (2012): regret in $\mathbb{R}$ that depends on $|u| \log |u|$ instead of $|u|^2 + 1$.
- Orabona (2013): generalization to Hilbert space.
- McMahan&Orabona (2014): $\|u\| \sqrt{\log(\|u\| + 1)}$ regret.
- Orabona (2014): link between new online algorithms and self-tuning SVMs, and a data dependent bound.
- A parallel line of work on adaptive learning with expert advice: Chaudhuri et al. (2009), Chernov&Vovk (2010), Luo&Schapire (2014, 2015), Koolen&van-Erven (2015), Foster et al. (2015).
- Orabona&Pál (2016): parameter-free algorithms for online learning from coin-betting.

PARAMETER-FREE SGD BASED ON THE KT ESTIMATOR

Require: Function $f(w, x, y)$ convex in w
Require: Training set $\{x_i, y_i\}_{i=1}^N$
Require: Desired number of iterations T
  Initialize $\text{Wealth}_0 \leftarrow 1$ and $\theta_0 \leftarrow 0$
  for $t = 1, 2, \ldots, T$ **do**
    Set $w_t \leftarrow \text{Wealth}_{t-1} \frac{\theta_{t-1}}{t}$
    Select an index j at random from $\{1, 2, \ldots, N\}$
    Update $\theta_t \leftarrow \theta_{t-1} - \nabla f(w_t, x_j, y_j)$
    $\text{Wealth}_t \leftarrow \text{Wealth}_{t-1} - \langle \nabla f(w_t, x_j, y_j), w_t \rangle$
  end for
  Output $\overline{w}_T = \frac{1}{T}\sum_{t=1}^T w_t$

THEORETICAL GUARANTEES

One epoch: $T \le N$
The average $\overline{w}_T$ is an approximate minimizer of the *risk* $\mathbf{E}[f(w, X, Y)]$:

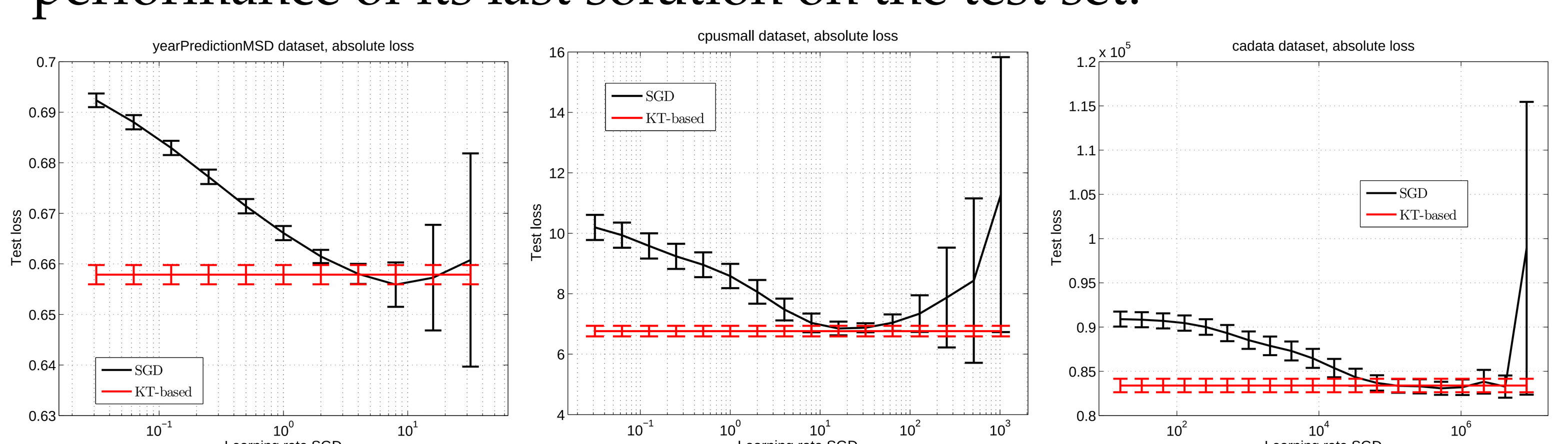
$$\mathbf{E}[f(\overline{w}_T, X, Y)] - \mathbf{E}[f(w^*, X, Y)] \le \frac{\|w^*\|}{\sqrt{T}}\sqrt{\log(1 + 4T^2\|w^*\|^2)} + \frac{1}{T}.$$

Multiple epochs: $T > N$
The average $\overline{w}_T$ is an approximate minimizer of the *training set error* $F(w) = \sum_{i=1}^N f(w, x_i, y_i)$:

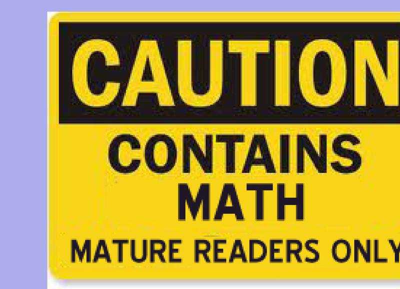
$$\mathbf{E}[F(\overline{w}_T)] - F(\widehat{w}) \le \frac{\|\widehat{w}\|}{\sqrt{T}}\sqrt{\log(1 + 4T^2\|\widehat{w}\|^2)} + \frac{1}{T}.$$

DOES IT WORK FOR REAL?

- Split data into 75% training + 25% test
- Train with one pass over the training set and evaluate the final classifier on the test set.
- Use 5 different splits into training+test. Report average and standard deviation.
- We have run SGD with different learning rates and shown the performance of its last solution on the test set.



- Clearly, the optimal learning rate of SGD is completely data-dependent.
- Interestingly, the performance of SGD becomes very unstable with large learning rates.
- Yet *our parameter-free algorithm has a performance very close to the unknown optimal tuning of the learning rate of SGD.*



LEARNING RATES IN ONLINE LINEAR LEARNING

- Define
$$\text{Regret}_T(u) = \sum_{t=1}^T \langle \ell_t, w_t \rangle - \sum_{t=1}^T \langle \ell_t, u \rangle.$$
- OGD with learning rate η satisfies
$$\forall u \in \mathcal{H} \qquad \text{Regret}_T(u) \le \frac{\|u\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|\ell_t\|^2.$$
- Optimal oracle choice: $\eta = \frac{\|u\|}{\sqrt{\sum_{t=1}^T \|\ell_t\|^2}}$.
- Many algorithms adapt to the norms of the gradients (e.g. AdaGrad) while neglecting dependency on $\|u\|$.
- Adapting to u is *more difficult and more important.*
- Better guarantees are indeed possible: Streeter&McMahan (2012), Orabona (2013), McMahan&Abernethy (2013), McMahan&Orabona (2014), Orabona (2014)
$$\forall u \in \mathcal{H} \qquad \text{Regret}_T(u) \le (O(1) + \text{polylog}(1 + \|u\|)\,\|u\|)\sqrt{T}.$$

REGRET GUARANTEE

Theorem. *Let $\{\ell_t\}_{t=1}^\infty$ be any sequence of loss vectors in a Hilbert space $\mathcal{H}$ such that $\|\ell_t\| \le 1$. The KT-based online algorithm satisfies*

$$\forall T \ge 0,\ \forall u \in \mathcal{H} \quad \text{Regret}_T(u) \le \|u\| \sqrt{T \ln\left(1 + 4T^2\|u\|^2\right)} + 1.$$

Proof Sketch.

- Duality between wealth and regret: Let $F : \mathcal{H} \to \mathbb{R}$ be convex. For any $w_1, \ldots, w_T$ and $g_1, \ldots, g_T$,
$$\underbrace{\sum_{t=1}^T \langle g_t, w_t \rangle}_{\text{Reward}_T} \ge F\left(\sum_{t=1}^T g_t\right) \Leftrightarrow \forall u \in \mathcal{H},\ \underbrace{\sum_{t=1}^T \langle g_t, u - w_t \rangle}_{\text{Regret}_T(u)} \le F^*(u).$$
- Consider the 1-dimensional case $\mathcal{H} = \mathbb{R}^1$.
- Set $w_t = \beta_t \text{Wealth}_{t-1}$ where β_t is the KT estimator.
- If $\ell_t \in \{+1, -1\}$, the results follows directly from the guarantee on the KT estimator and duality above.
- Extend to $\ell_t \in [-1, 1]$ by convexity: worst ℓ_t is in $\{+1, -1\}$.
- Extend 1-d case to Hilbert space: Worst direction of ℓ_t is the same as the direction of $\sum_{s=1}^{t-1} \ell_s$.