A SGD safari

Lorenzo Rosasco
University of Genova
Massachusetts Institute of Technology - Istituto Italiano di Tecnologia
lcsl.mit.edu

Jan. 3rd, 2019 – DALI 2019: Optimization Workshop

joint work with R. Camoriano (LCSL), J. Lin (EPFL), S. Villa (UniGE)

Outline

Classic results

Statistical learning & least squares

Multi-pass SGD

SGD

Problem Solve

$$\min_w F(w), \qquad F(w) = \mathbb{E}_Z \ell(Z, w)$$

SGD

$$w_{t+1} = w_t - \gamma_t \nabla \ell(Z_t, w_t), \qquad t = 0, \dots, T$$

- It holds $\mathbb{E}\nabla L(Z_t, w) = \nabla F(w)$, hence the name.
- Every step requires a new gradient estimates.

SGD typical result

Assume F convex, smooth, with bounded gradients and take $\gamma \lesssim \frac{1}{\sqrt{t}}$, then

$$\mathbb{E}\left[F(w_T) - \min_w F(w)\right] \lesssim \frac{1}{\sqrt{T}}.$$

- Rates are optimal improved.
- Better rates under stronger conditions: strong convexity, KL/conditioning

SGD for training error

Special case

$$F(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, w),$$

Z rand. var. uniformly distributed on $z_1, \ldots, z_n$.

- Better rates achievable in this case.
- Again improvable under stronger conditions: strong convexity, KL/conditioning.
- SGD called also incremental gradient in this case.

Understanding SGD: from practice to theory

- **multiple-passes (gradients are re-used)**
- **various step-size choices**
- **mini-batch**
- averaging
- sketching
- acceleration
- preconditioning
- ...

What is the impact for learning (test error)?

Outline

Classic results

Statistical learning & least squares

Multi-pass SGD

Least squares learning

- $\mathcal{X}$ Hilbert space
- $Z = (X, Y)$ with values in $\mathcal{X} \times \mathbb{R}$

Problem:
Solve

$$\min_{w \in \mathcal{X}} \mathcal{E}(w) \qquad \mathcal{E}(w) = \mathbb{E}[(Y - \langle w, X \rangle)^2]$$

given only $(x_i, y_i)_{i=1}^n$ i.i.d.

Minimal norm solution:

$$w^\dagger = \operatorname*{argmin}_{w \in \mathcal{O}} \|w\|, \qquad \mathcal{O} = \operatorname*{argmin}_{w \in \mathcal{X}} \mathcal{E}(w)$$

Ill-posedeness

Least squares optimality conditions:

$$\Sigma w^\dagger = g, \qquad \Sigma = \mathbb{E}[X \otimes X], \quad g = \mathbb{E}[XY]$$

and $w^\dagger \in \mathsf{Null}(\Sigma)^\perp$

Ill-posedeness

Least squares optimality conditions:

$$\Sigma w^\dagger = g, \qquad \Sigma = \mathbb{E}[X \otimes X], \quad g = \mathbb{E}[XY]$$

and $w^\dagger \in \mathsf{Null}(\Sigma)^\perp$

Ill-posedness

- $\mathcal{X}$ **infinite dimensional,** Σ **compact** $\Rightarrow$ **problem is ill-posed.**
- if $\mathcal{X}$ is finite dimensional it is well posed, but possibly badly conditioned.

Least squares SGD

$$\widehat{w}_{t+1} = \widehat{w}_t - \eta_t(x_{i_t}(\langle \widehat{w}_t, x_{i_t} \rangle - y_{i_t}) + \lambda \widehat{w}_t), \quad t = 0, \dots T$$

Free parameters:

- regularization parameter λ
- step-size $(\eta_t)_t$
- stopping time T, ($T > n$ multiple "passes")

Note: $(i_t)_t$ deterministic or stochastic (with/without replacement)

LS-SGD: Previous results

Non asymptotic:

- [Smale-Yao '05] Fixed λ (some classic results hold for this case).
- [Tarres-Yao '07] Decreasing λ.
- [Ying-Pontil '07] $\lambda = 0$.

All one pass, i.e. $i_t = t$, and with decreasing step-size.

[Villa-Rosasco '15] $\lambda = 0$, **multiple passes** (for the first time?), cyclic selection.

Outline

Classic results

Statistical learning & least squares

Multi-pass SGD

Multi-pass LS-SGD

$$\widehat{w}_{t+1} = \widehat{w}_t - \eta(x_{i_t}(\langle \widehat{w}_t, x_{i_t} \rangle - y_{i_t})), \quad t = 0, \dots T$$

Note: $(i_t)_t$ chosen **uniformly at random with replacement**

Multi-pass LS-SGD

$$\widehat{w}_{t+1} = \widehat{w}_t - \eta(x_{i_t}(\langle \widehat{w}_t, x_{i_t} \rangle - y_{i_t})), \quad t = 0, \dots T$$

Note: $(i_t)_t$ chosen **uniformly at random with replacement**

Theorem (Lin, R. '16)

Assume $\|X\| \leq 1$ *and* $|Y| \leq 1$ *for all* η *and* T*,*

$$\mathbb{E}\mathcal{E}(\widehat{w}_T) - \mathcal{E}(w^\dagger) \lesssim \frac{1}{\eta T} + \frac{1}{\sqrt{n}} \left(\frac{\eta T}{\sqrt{n}} \right)^2 + \eta \left(1 \vee \frac{\eta T}{\sqrt{n}} \right)$$

Multi-pass LS-SGD

$$\widehat{w}_{t+1} = \widehat{w}_t - \eta(x_{i_t}(\langle \widehat{w}_t, x_{i_t} \rangle - y_{i_t})), \quad t = 0, \dots T$$

Note: $(i_t)_t$ chosen **uniformly at random with replacement**

Theorem (Lin, R. '16)

Assume $\|X\| \leq 1$ *and* $|Y| \leq 1$ *for all* η *and* T*,*

$$\mathbb{E}\mathcal{E}(\widehat{w}_T) - \mathcal{E}(w^\dagger) \lesssim \frac{1}{\eta T} + \frac{1}{\sqrt{n}} \left(\frac{\eta T}{\sqrt{n}} \right)^2 + \eta \left(1 \vee \frac{\eta T}{\sqrt{n}} \right)$$

Note

- **Statistics and optimization:** integrated in the bound.
- **Bias-variance**: parameter choices derived optimizing the bound.

Multi-pass vs one pass SGD

Corollary (Lin, R. '16)

Assume $\|X\| \leq 1$ *and* $|Y| \leq 1$ *a.s. and let*

- $T = n$ *(1 pass),* $\eta = \frac{1}{\sqrt{n}}$.
- $T = n^{3/2}$ *(*$\sqrt{n}$ *passes),* $\eta = \frac{1}{n}$.

Then,

$$\mathbb{E}\mathcal{E}(\widehat{w}_T) - \mathcal{E}(w^\dagger) \lesssim \frac{1}{\sqrt{n}}$$

Note

- Optimal (nonparametric) rate in a minmax sense.
- With a larger step-size, one pass suffices (recovering [Dieulevet, Bach '14– Ying, Pontil, '06]).

Beyond the worst case: source condition

Recall

$$\Sigma w^\dagger = g, \qquad \Sigma = \mathbb{E}[X \otimes X], \quad g = \mathbb{E}[XY]$$

and $w^\dagger \in \mathsf{Null}(\Sigma)^\perp$

- **S) Source condition** $w^\dagger \in \mathsf{Range}\,(\Sigma^\alpha)$, $\alpha > 0$
- **C) Capacity condition** $\sigma_i(\Sigma) \sim i^{-\gamma}, \quad \gamma \in (0, 1]$

Fast rates

Theorem (Lin, R. '16)

Assume $\|X\| \leq 1$, $|Y| \leq 1$ *and* $S), C)$ *hold. Then, for all* η *and* T,

$$\mathbb{E}\mathcal{E}(\widehat{w}_T) - \mathcal{E}(w^\dagger) \lesssim \left(\frac{1}{\eta T}\right)^{2\alpha+1} + \frac{1}{n^{\frac{2\alpha+1}{2\alpha+1+\gamma}}} \left(\frac{\eta T}{n^{\frac{1}{2\alpha+1+\gamma}}}\right)^2 + \eta \left(1 \vee \frac{\eta T}{n^{\frac{1}{2\alpha+1+\gamma}}}\right)$$

Note

- Reduces to worst case for $\alpha = 0, \quad \gamma = 1$.
- Different **parameter choices** derived optimizing the bound.

Multiple passes SGD

Corollary (Lin, R. '16)

Assume $\|X\| \leq 1$, $|Y| \leq 1$ *and* $S), C)$ *hold. Let*

- $T = n^{\frac{1}{2\alpha+1+\gamma}+1}$ *(*$n^{\frac{1}{2\alpha+1+\gamma}}$ *passes)*
- $\eta = \frac{1}{n}$.

Then,

$$\mathbb{E}\mathcal{E}(\widehat{w}_T) - \mathcal{E}(w^\dagger) \lesssim n^{-\frac{2\alpha+1}{2\alpha+1+\gamma}}$$

Note

- Optimal (nonparametric) rate in a minmax sense.
- Same as Tikhonov regularization but include optimization!
- Choosing T_n by cross validation (CV) achieves the same rate.

One pass SGD

Corollary (Dieulevet, Bach '16)

Assume $\|x\| \leq 1$, $|y| \leq 1$ *and* $S), C)$ *hold with* $\alpha < 1/2$. *Let*

- $T = n$ *(1 pass)*
- $\eta = n^{-\frac{2\alpha+1}{2\alpha+1+\gamma}}$.
- $\bar{w}_n = \frac{1}{n}\sum_{t=1}^{n} w_t$.

Then,

$$\mathbb{E}\mathcal{E}(\bar{w}_T) - \mathcal{E}(w^\dagger) \lesssim n^{-\frac{2\alpha+1}{2\alpha+1+\gamma}}$$

Note

- Optimal (nonparametric) rate in a minmax sense.
- Same rates using cross validation (CV) for choosing step-size η.

Remarks

- Stepsize and iterations control convergence and stability of SGD: one of the two (or both) needs be **tuned**.
- Proof extends to **harder** **or** **easier** learning problems with slightly different take home messages [Pillaud et al. 1'8].
- Proof strategy extends to averaging [Pillaud et al. '18], decaying stepsize, **mini-batches** [Lin, R.'16].

Mini-batch, multi-pass LS-SGD

$$\widehat{w}_{t+1} = \widehat{w}_t - \eta_t \frac{1}{b} \sum_{i=b(t-1)+1}^{bt} \big(\langle \widehat{w}_t, (x_{j_i})\rangle - y_{j_i}\big)(x_{j_i})$$

Theorem (Lin, R. '16)

Assume $\|X\| \le 1$ *and* $|Y| \le 1$ *for all* η *and* T*,*

$$\mathbb{E}\mathcal{E}(\widehat{w}_T) - \mathcal{E}(w^\dagger) \lesssim \frac{1}{\eta T} + \frac{1}{\sqrt{n}} \left(\frac{\eta T}{\sqrt{n}}\right)^2 + \frac{\eta}{b}\left(1 + \frac{\eta T}{\sqrt{n}}\right)$$

Note

- **mini-batch size:** b.

Multi-pass vs one pass SGD

Corollary (Lin, R. '16)

Assume $\|X\| \leq 1$ *and* $|Y| \leq 1$ *a.s. and consider one of the following choices*

1. $b = 1$, $\eta_t \simeq \frac{1}{\sqrt{n}}$, *and* $T = n$ *iterations (*1 *pass over the data);*
2. $b = \sqrt{n}$, $\eta_t \simeq 1$, *and* $T = \sqrt{n}$ *iterations (*1 *pass over the data);*
3. $b = n$, $\eta_t \simeq 1$, *and* $T = \sqrt{n}$ *iterations (*$\sqrt{n}$ *passes over the data);*

Then,

$$\mathbb{E}\mathcal{E}(\widehat{w}_T) - \mathcal{E}(w^\dagger) \lesssim \frac{1}{\sqrt{n}}$$

Note

- Mini-batching allows larger step-sizes.
- No gain after $b = \sqrt{n}$.
- Refined results beyond this worst case.

Concluding

- Tools from statistical learning to understand practically used SGD.
- First optimal results for multiple passes (and minibatching).
- Sketching/random features $\rightarrow$ I brought a poster...

Some open problems

- Combine averaging and minibatching - [Mücke, R. '19] on the way
- Beyond least squares – [Hardt et al. 16, Lin, Camoriano R. '16] partial results
- Beyond minimal ℓ_2 norm – [Matet, R., Villa, Vu '16, Garrigos, R., Villa '16] batch case
- Acceleration - results in [Jain et al '16-]
- Non-convexity

References

- Learning with incremental iterative regularization
  L Rosasco, S Villa
  Advances in Neural Information Processing Systems, 1630-1638
- Optimal rates for multi-pass stochastic gradient methods
  J Lin, L Rosasco
  The Journal of Machine Learning Research 18 (1), 3375-3421
- Generalization properties and implicit regularization for multiple passes SGM
  J Lin, R Camoriano, L Rosasco
  International Conference on Machine Learning, 2340-2348
- Learning with sgd and random features
  L Carratino, A Rudi, L Rosasco
  Advances in Neural Information Processing Systems, 10213-10224