From inexact optimization to learning via gradient concentration

Bernhard Stankewitz
AIP 2023, Göttingen

Department of Decision Sciences
Bocconi University

Collaborators

Nicole Mücke, Braunschweig

Lorenzo Rosasco, Genova

Setting

- $(\mathcal{H}, \|\cdot\|)$ is a real, separable Hilbert space and $\mathcal{Y} \subset \mathbb{R}$. Consider n i.i.d. observations of $(X, Y) \in \mathcal{H} \times \mathcal{Y}$, with $\|X\| \leq \kappa \in [1, \infty)$ **(Bound)**.
- Learn linear relationship between X and Y expressed as $w \in \mathcal{H}$. Suffer the loss $\ell(Y, \langle X, w\rangle)$, which is convex **(Conv)**, Lipschitz with constant $L > 0$ **(Lip)** and has Lipschitz gradients with constant $M > 0$ **(Smooth)**. Includes classical kernel learning, see e.g. Rosasco and Villa [RV15].
- Minimize the population risk $\mathcal{L}(w) := \mathbb{E}_{(X,Y)}[\ell(Y, \langle X, w\rangle)]$ with minimizer $w_* \in \mathcal{H}$ **(Min)**.

GD-Algorithm

1. Set $v_0 = 0 \in \mathcal{H}$ and choose constant stepsize $\gamma > 0$.
2. For $t = 0, 1, 2, \ldots,$ define the GD-iteration

$$v_{t+1} = v_t - \gamma \nabla \widehat{\mathcal{L}}(v_t) = v_t - \frac{\gamma}{n} \sum_{j=1}^{n} \ell'(Y_j, \langle X_j, v_t\rangle) X_j.$$

3. For some $T \geq 1$, choose the last iterate v_T or the averaged iterate $\overline{v}_T := T^{-1} \sum_{t=1}^{T} v_t$.

Excess risk decomposition

Proposition (Excess risk decomposition)

Suppose assumptions **(Bound)**, **(Conv)**, **(Smooth)** *and* **(Min)** *are satisfied. Consider the GD-iteration with constant step size* $\gamma \leq 1/(\kappa^2 M)$*. Then, the excess risk of the averaged iterate* $\overline{v}_T$ *satisfies*

$$\mathcal{L}(\overline{v}_T) - \mathcal{L}(w_*) \leq \frac{\|w_*\|^2}{2\gamma T} + \frac{1}{T}\sum_{t=1}^{T}\langle \nabla\mathcal{L}(v_{t-1}) - \nabla\widehat{\mathcal{L}}(v_{t-1}), v_t - w_*\rangle.$$

- Inspired by the literature on *inexact optimization* ("$\nabla\mathcal{L}(v_t) + e_t$"), see e.g. Bertsekas and Tsitsiklis [BT00], Schmidt, Roux, and Bach [SRB11] and Yang, Wei, and Wainwright [YWW19].[1]
- Recovers the deterministic optimization setting, see Bubeck [Bub15].
- In order to bound the stochastic error, it is sufficient to solve two problems:
  1. Bound the gradient path $(v_t)_{t\geq 0}$ in a ball with radius $R > 0$.
  2. Bound the empirical process $\sup_{\|v\|\leq R} \|\nabla\widehat{\mathcal{L}}(v) - \nabla\mathcal{L}\|$.

[1] F. Yang, Y. Wei, M. Wainwright. "Early stopping for kernel boosting algorithms: A general analysis with localized complexities" In: *IEEE Transactions on Information Theory* (2019).

Comparison with the classical approach

Classical decomposition.

$$\mathcal{L}(v_t) - \mathcal{L}(w_*) = \underbrace{\mathcal{L}(v_t) - \widehat{\mathcal{L}}(v_t)}_{\text{(I)}} + \underbrace{\widehat{\mathcal{L}}(v_t) - \widehat{\mathcal{L}}(w_*)}_{\text{(II)}} + \underbrace{\widehat{\mathcal{L}}(w_*) - \mathcal{L}(w_*)}_{\text{(III)}} \tag{1}$$

- (II) is an optimization error treated by deterministic results.
- (I) and (III) treated by concentration for $\sup_{\|v\| \leq R} |\widehat{\mathcal{L}}(v) - \mathcal{L}(v)|$.
- Guarantee of the form $\|v_t\| \leq R$ is fundamental.
- Projected gradients Holland and Ikeda [HI18], clipped gradients Gorbunov, Danilova, and Gasnikov [GDG20], technical analyses in Lin, Rosasco, and Zhou [LRZ16], Lei, Hu, and Tang [LHT21].

Gradient concentration

Proposition (Gradient concentration)

Suppose assumptions **(Bound)**, **(Lip)** *and* **(Smooth)** *are satisfied and let* $R > 0$*. Then, for any* $\delta > 0$*, with probability at least* $1 - \delta$*,*

$$\sup_{\|v\| \leq R} \|\nabla \mathcal{L}(v) - \nabla \widehat{\mathcal{L}}(v)\| \leq 4\widehat{\mathcal{R}}_n(\nabla \ell \circ \mathcal{F}_R) + \kappa L \sqrt{\frac{2\log(4/\delta)}{n}} + \kappa L \frac{4\log(4/\delta)}{n}.$$

- The empirical Rademacher complexity can be bounded by

$$\widehat{\mathcal{R}}(\nabla \ell \circ \mathcal{F}_R) \leq \frac{2\sqrt{2}(\kappa L + \kappa^2 M R)}{\sqrt{n}}. \tag{2}$$

- Concentration arguments from Foster, Sekhari, and Sridharan [FSS18][2], vector contraction inequality from Maurer [Mau16].[3]

[2] D. J. Foster, A. Sekhari and K. Sridharan. "Uniform convergence of gradients for non-convex learning and optimization". In: *Advances in Neural Information Processing Systems* 2018.

[3] A. Maurer. " A vector-contraction inequality for Rademacher complexities". In *Algorithmic Learning Theory* 9925 (2016), pp. 3-17.

Bounding the gradient path

For $e_s = \nabla\widehat{\mathcal{L}}(v_s) - \nabla\mathcal{L}(v_s)$, inductively control $\|v_t\|$ via

$$\|v_{t+1} - w_*\|^2 \leq \|v_0 - w_*\|^2 + 2\sum_{s=0}^{t} \gamma\big(\langle -e_s, v_s - w_*\rangle + \kappa L\|e_s\|\big). \tag{3}$$

Proposition (Bounded gradient path)

Suppose assumptions **(Bound)**, **(Conv)**, **(Lip)**, **(Smooth)** *and* **(Min)** *are satisfied, set* $v_0 = 0$ *and choose a constant step size* $\gamma \leq \min\{1/(\kappa^2 M), 1\}$. *Fix* $\delta \in (0, 1]$ *with*

$$\sqrt{n} \geq \max\{1, 90\gamma T\kappa^2(1 + \kappa L)(M + L)\}\sqrt{\log(4/\delta)} \tag{4}$$

and $R = \max\{1, 3\|w_*\|\}$. *Then, on the gradient concentration event from Proposition 0.2 with probability at least* $1 - \delta$ *for the above choice of* R, *we have*

$$\|v_t\| \leq R \qquad \text{and} \qquad \|v_t - w_*\| \leq \frac{2R}{3} \qquad \text{for all } t = 1, \ldots, T.$$

Combining the ingredients

Combine the excess risk bound

$$\mathcal{L}(\overline{v}_T) - \mathcal{L}(w_*) \leq \frac{\|w_*\|^2}{2\gamma T} + \frac{1}{T}\sum_{t=1}^{T}\langle \nabla\mathcal{L}(v_{t-1}) - \nabla\widehat{\mathcal{L}}(v_{t-1}), v_t - w_*\rangle. \tag{5}$$

with gradient concentration and the bounded gradient path:

- $\sup_{\|v\|\leq R} \|\nabla\mathcal{L}(v) - \nabla\widehat{\mathcal{L}}(c)\| \lesssim R\sqrt{\log(4/\delta)/n}$ with probability at least $1-\delta$.
- $\|v_t\| \leq R \sim \|w_*\|$, $t \leq T$ for n large enough on the same event.

Excess risk bounds

Theorem (Excess Risk bound, averaged iterate)

Suppose Assumptions **(Bound)**, **(Conv)**, **(Lip)**, **(Smooth)** *and* **(Min)** *are satisfied, set* $v_0 = 0$ *and choose a constant step size* $\gamma \leq \min\{1/(\kappa^2 M), 1\}$ *in the GD-iteration. Then, for any* $\delta \in (0, 1]$, *such that*

$$\sqrt{n} \geq \max\{1, 90\gamma T \kappa^2 (1 + \kappa L)(M + L)\} \sqrt{\log(4/\delta)},$$

the averaged iterate $\overline{v}_T$ *satisfies that, with probability at least* $1 - \delta$,

$$\mathcal{L}(\overline{v}_T) - \mathcal{L}(w_*) \leq \frac{\|w_*\|^2}{2\gamma T} + 180 \max\{1, \|w_*\|^2\} \kappa^2 (M + L) \sqrt{\frac{\log(4/\delta)}{n}}.$$

Setting $\gamma T = \sqrt{n}/(90\kappa^2(1 + \kappa L)(M + L)\sqrt{\log(4/\delta)})$ *yields*

$$\mathcal{L}(\overline{v}_T) - \mathcal{L}(w_*) \leq 225 \max\{1, \|w_*\|^2\} \kappa^2 (1 + \kappa L)(M + L) \sqrt{\frac{\log(4/\delta)}{n}}.$$

Excess risk bounds

Theorem (Excess Risk bound, last iterate)

Suppose Assumptions **(Bound)**, **(Conv)**, **(Lip)**, **(Smooth)** *and* **(Min)** *are satisfied, set* $v_0 = 0$ *and choose a constant step size* $\gamma \leq \min\{1/(\kappa^2 M), 1\}$ *in the GD-iteration. Then, for any* $\delta \in (0, 1]$, *such that*

$$\sqrt{n} \geq \max\{1, 90\gamma T \kappa^2 (1 + \kappa L)(M + L)\} \sqrt{\log(4/\delta)},$$

the last iterate v_T *satisfies that, with probability at least* $1 - \delta$,

$$\mathcal{L}(\overline{v}_T) - \mathcal{L}(w_*) \leq \frac{\|w_*\|^2}{2\gamma T} + 425 \max\{1, \|w_*\|^2\} \kappa^2 (M + L) \sqrt{\frac{\log(4/\delta)}{n}}.$$

Setting $\gamma T = \sqrt{n}/(90\kappa^2(1 + \kappa L)(M + L)\sqrt{\log(4/\delta)})$ *yields*

$$\mathcal{L}(\overline{v}_T) - \mathcal{L}(w_*) \leq 470 \max\{1, \|w_*\|^2\} \kappa^2 (1 + \kappa L)(M + L) \sqrt{\frac{\log(4/\delta)}{n}}.$$

Simulation example

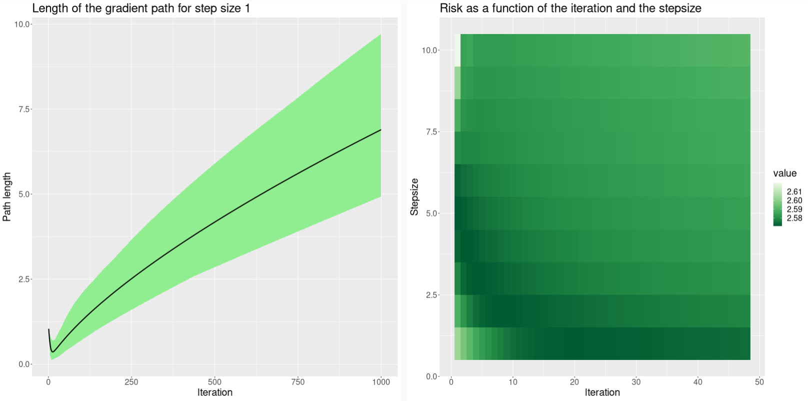

Figure 1: Simulation results for the logistic loss.

Thank you!

References

[BT00] D. P. Bertsekas and J. N. Tsitsiklis. **"Gradient Convergence in Gradient Methods with Errors".** In: *SIAM Journal on Optimization* 10.3 (2000), pp. 627–642.

[Bub15] S. Bubeck. **"Convex Optimization: Algorithms and Complexity".** In: *Foundations and Trends in Machine Learning* 8.3–4 (2015), pp. 231–357.

[FSS18] D. J. Foster, A. Sekhari, and K. Sridharan. **"Uniform Convergence of Gradients for Non-convex Learning and Optimization".** In: *Advances in Neural Information Processing Systems*. Vol. 31. Redhook, NY: Curran Associates, Inc., 2018.

[GDG20] E. Gorbunov, M. Danilova, and A. Gasnikov. **"Stochastic Optimization with Heavy-tailed Noise via Accelerated Gradient Clipping".** In: *Advances in Neural Information Processing Systems*. Redhook, NY: Curran Associates, Inc., 2020.

[HI18] M. J. Holland and K. Ikeda. **Efficient Learning with Robust Gradient Descent.** 2018. URL: https://arxiv.org/abs/1706.00182.

[LHT21] Y. Lei, T. Hu, and K. Tang. **“Generalization Performance of Multi-pass Stochastic Gradient Descent with Convex Loss Functions”.** In: *The Journal of Machine Learning Research* 22.25 (2021), pp. 1–41.

[LRZ16] J. Lin, L. Rosasco, and D. Zhou. **“Iterative Regularization for Learning with Convex Loss Functions”.** In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 2718–2755.

[Mau16] A. Maurer. **“A Vector-contraction Inequality for Rademacher Complexities”.** In: *Algorithmic Learning Theory*. Vol. 9925.: Springer, 2016, pp. 3–17.

[RV15] L. Rosasco and S. Villa. **“Learning with Incremental Iterative Regularization”.** In: *Advances in Neural Information Processing Systems*. Vol. 28. Redhook, NY: Curran Associates, Inc., 2015.

[SRB11] M. Schmidt, N. Roux, and F. Bach. **“Convergence Rates of Inexact Proximal-gradient Methods for Convex Optimization”.** In: *Advances in Neural Information Processing Systems*. Vol. 24. Redhook, NY: Curran Associates, Inc., 2011.

[YWW19] F. Yang, Y. Wei, and M. J. Wainwright. **"Early stopping for kernel boosting algorithms: A general analysis with localized complexities".** In: *IEEE Transactions on Information Theory* (2019).