Hyperparameter Tuning via Trajectory Predictions: Stochastic Prox-Linear Methods in Matrix Sensing

Author(s):
Lou, Mengqi; Verchand, Kabir Aladin; Pananjady, Ashwin

Publication date:
2024-03-06

Permanent link:
https://doi.org/10.3929/ethz-b-000664567

Hyperparameter tuning via trajectory predictions: Stochastic prox-linear methods in matrix sensing

Mengqi Lou[†], Kabir Aladin Verchand[*], and Ashwin Pananjady[†‡]
[*]Statistical Laboratory, University of Cambridge
Cambridge, UK
email: kav29@cam.ac.uk
[†]Georgia Institute of Technology, School of Industrial and Systems Engineering, {mlou30, ashwinpm}@gatech.edu
[‡]Georgia Institute of Technology, School of Electrical and Computer Engineering, ashwinpm@gatech.edu

We consider estimating a rank one matrix $\boldsymbol{\mu}_\star\boldsymbol{\nu}_\star^\top \in \mathbb{R}^{d\times d}$ from i.i.d. observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ drawn in an online, mini-batched fashion according to the model $y_i = \langle \boldsymbol{x}_i, \boldsymbol{\mu}_\star\rangle \cdot \langle \boldsymbol{z}_i, \boldsymbol{\nu}_\star\rangle + \epsilon_i$. To do so, we consider minimizing the population loss corresponding to the negative log-likelihood, namely $\bar{L}(\boldsymbol{\mu}, \boldsymbol{\nu}) = \mathbb{E}\{\big(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\mu}\rangle \cdot \langle \boldsymbol{z}_i, \boldsymbol{\nu}\rangle\big)^2\}$, which we emphasize is a non-convex function of the inputs. Towards minimizing the population loss, consider an iterate $(\boldsymbol{\mu}_t, \boldsymbol{\nu}_t)$. We take mini-batches of size m with samples $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)_{i=1}^m$ and form the data $\boldsymbol{a}_i^\top = \begin{bmatrix}\boldsymbol{z}_i^\top \boldsymbol{\nu}_t \boldsymbol{x}_i^\top & \boldsymbol{x}_i^\top \boldsymbol{\mu}_t \boldsymbol{z}_i^\top\end{bmatrix}$ for each $1 \le i \le m$; define the pair of diagonal matrices $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{X}\boldsymbol{\mu}_t)$, $\widetilde{\boldsymbol{W}} = \mathrm{diag}(\boldsymbol{Z}\boldsymbol{\nu}_t)$; and collect the vectors $\boldsymbol{a}_i$ into a concatenated data matrix $\boldsymbol{A} = [\boldsymbol{a}_1 \mid \boldsymbol{a}_2 \mid \ldots \mid \boldsymbol{a}_m]^\top = \begin{bmatrix}\widetilde{\boldsymbol{W}}\boldsymbol{X} \mid \boldsymbol{W}\boldsymbol{Z}\end{bmatrix} \in \mathbb{R}^{m\times 2d}$. We then consider the following stochastic prox-linear update to define the next iterate $(\boldsymbol{\mu}_{t+1}, \boldsymbol{\nu}_{t+1})$

$$\begin{bmatrix}\boldsymbol{\mu}_{t+1}\\ \boldsymbol{\nu}_{t+1}\end{bmatrix} = \boldsymbol{A}_\lambda^{-1}\Big(\boldsymbol{A}^\top\big(\boldsymbol{y} + \mathrm{diag}(\boldsymbol{W}\widetilde{\boldsymbol{W}})\big) + \lambda m \begin{bmatrix}\boldsymbol{\mu}_t\\ \boldsymbol{\nu}_t\end{bmatrix}\Big),$$

where λ denotes an inverse step-size parameter and $\boldsymbol{A}_\lambda = \boldsymbol{A}^\top\boldsymbol{A} + \lambda m \boldsymbol{I}$. Our main contribution is to provide a deterministic prediction of the trajectory of the iterative method defined in the previous display under the pair of assumptions $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}_{i\ge 1} \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \boldsymbol{I}_d)$ and $\|\boldsymbol{\mu}_\star\|_2 = \|\boldsymbol{\nu}_\star\|_2 = 1$. More concretely, we obtain the following.

a) Sharp, deterministic predictions which adapt to problem error: Consider running one-step of the prox-linear update starting from a pair $(\boldsymbol{\mu}_\sharp, \boldsymbol{\nu}_\sharp)$ and let $[\boldsymbol{\mu}_+^\top \mid \boldsymbol{\nu}_+^\top]^\top$ denote the next iterate. For *all* minibatch sizes $1 \le m \le d$ and a large range of step-sizes $\lambda \gtrsim (1+\sigma)d/m$, we derive an explicit, deterministic, four-dimensional prediction that closely tracks the error of its empirical counterparts. We additionally prove a non-asymptotic guarantee on our predictions, showing that its fluctuations scale as $\frac{\|\boldsymbol{\mu}_\sharp\boldsymbol{\nu}_\sharp^\top - \boldsymbol{\mu}_\star\boldsymbol{\nu}_\star^\top\|_F + \sigma}{\lambda\sqrt{m}}$, up to poly-logarithmic in dimension factors. Note that this guarantee—in contrast to previous work [1], [2]—provides bounds on the deviation which scale with the *current* estimation error $\|\boldsymbol{\mu}_\sharp\boldsymbol{\nu}_\sharp^\top - \boldsymbol{\mu}_\star\boldsymbol{\nu}_\star^\top\|_F$. This, in turn, enables a transparent convergence analysis of the iterations for all noise levels $\sigma \ge 0$.

Our proof reposes on a variant of El Karoui, et. al's leave-one-out method [3]. In particular, given the ground truth $\boldsymbol{\mu}_\star$ and a current iterate $\boldsymbol{\mu}_\sharp$, we let $\mathbb{U} = \{\boldsymbol{\mu}_\star, \boldsymbol{P}_{\boldsymbol{\mu}_\star}^\perp\boldsymbol{\mu}_\sharp/\|\boldsymbol{P}_{\boldsymbol{\mu}_\star}^\perp\boldsymbol{\mu}_\sharp\|_2, \boldsymbol{u}_3, \ldots, \boldsymbol{u}_d\}$ denote an orthonormal basis of $\mathbb{R}^d$. We obtain a closed form expression for each of the projections $\langle \boldsymbol{\mu}_+, \boldsymbol{u}\rangle$, $\boldsymbol{u} \in \mathbb{U}$. We then use standard tools in random matrix theory to obtain deterministic predictions of each of these projections.

b) Fine-grained convergence analysis: We use our deterministic predictions to execute an iterate-by-iterate analysis of the stochastic prox-linear algorithm from a local initialization. This analysis reveals several fine-grained properties of the convergence behavior. In particular, for the step-size choice $\lambda^{-1} \asymp m/(d(1+\sigma^2))$ and batch size $m \gtrsim \mathsf{polylog}(d)$, we show that it takes $\tau = \Theta\Big(\frac{d(1+\sigma^2)}{m} \cdot \log\Big(\frac{1}{\sigma^2}\Big)\Big)$ many iterations in order to guarantee an error $\|\boldsymbol{\mu}_\tau\boldsymbol{\nu}_\tau^\top - \boldsymbol{\mu}_\star\boldsymbol{\nu}_\star^\top\|_F^2 \lesssim \sigma^2$. This reveals a linear speed-up in the batch size m for *all* noise levels $\sigma \ge 0$. As a consequence, the total sample complexity for reaching estimation error σ^2 is $O(d(1+\sigma^2)\log(1/\sigma^2))$. Moreover, for other step-size choices $\lambda^{-1} \lesssim m/(d(1+\sigma^2))$, we show that it takes $\tau = \Theta\Big(\lambda \cdot \log\Big(\frac{\lambda m}{d\sigma^2}\Big)\Big)$ many iterations to guarantee an error $\|\boldsymbol{\mu}_\tau\boldsymbol{\nu}_\tau^\top - \boldsymbol{\mu}_\star\boldsymbol{\nu}_\star^\top\|_F^2 \lesssim \frac{\sigma^2 d}{\lambda m}$, which in turn quantifies the dependence of the convergence behavior on the step-size λ^{-1}. That is, decreasing the step-size λ^{-1} introduces a tension between the increasing iteration complexity and decreasing eventual estimation error. Note that our guarantees on iteration complexity are sharp in the sense that our bounds provide both upper and lower bounds on the rate of convergence.

Our convergence proofs rely on properties of the deterministic predictions. In particular, we first prove that the deterministic predictions enjoy sharp linear convergence. We then apply the deviation bounds on the deterministic predictions to transfer this property to the empirical iterates.

REFERENCES

[1] K.A. Chandrasekher, M. Lou, A. Pananjady, "Alternating minimization for generalized rank one matrix sensing: Sharp predictions from a random initialization," to appear in *Algorithmic Learning Theory*, San Diego, CA, 2024.

[2] K.A. Chandrasekher, A. Pananjady, C. Thrampoulidis, "Sharp global convergence guarantees for iterative nonconvex optimization with random data," *Annals of Statistics*, vol. 51, pp. 179–210, 2023.

[3] N. El Karoui, D. Bean, P.J. Bickel, C. Lim, B. Yu, "On robust regression with high-dimensional predictors," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 14557–14562, 2013.