Ergodic bilevel optimization

Christoph Brauer
Dirk A. Lorenz
Institute of Analysis and Algebra, TU Braunschweig
38092 Braunschweig, Germany
Email: {ch.brauer, d.lorenz}@tu-braunschweig.de

***Abstract*—In image and signal processing and beyond, quantities of interest are often reconstructed from noisy measurements by means of suitable convex optimization problems. While model based approaches usually assume that the ingredients of the problem are a priori known, data driven approaches are motivated by situations where the objective function or constraints of the problem are partially unknown and shall be learned from data. This gives rise to bilevel optimization problems in which the original convex problem (hereafter referred to as the lower-level problem) appears as a constraint. Applying gradient based algorithms to bilevel optimization problems poses the difficulty to differentiate through the solution operator of the lower-level problem. In this contribution, we consider the approach to unroll a fixed number of update steps of the Chambolle-Pock algorithm applied to the lower-level problem in order to accomplish this kind of differentiation approximately. We investigate the asymptotic behavior of the resulting gradients and conclude that unrolling ergodic averages instead of ordinary iterates can have a positive effect upon the learning dynamics.**

I. Introduction

We consider the task to recover a ground truth $\boldsymbol{x}^\dagger \in \mathbb{R}^n$ from degraded measurements $\tilde{\boldsymbol{x}} \in \mathbb{R}^n$. To that end, we assume that there exists an *unknown* analysis operator $\boldsymbol{K} \in \mathbb{R}^{m\times n}$ such that with

$$S(\boldsymbol{K}, \tilde{\boldsymbol{x}}) \coloneqq \operatorname*{argmin}_{\boldsymbol{x}\in\mathbb{R}^n} \; F(\boldsymbol{K}\boldsymbol{x}) \; + \; G(\boldsymbol{x} - \tilde{\boldsymbol{x}}) \tag{1}$$

$\hat{\boldsymbol{x}} \in S(\boldsymbol{K}, \tilde{\boldsymbol{x}})$ is a good approxmation of $\boldsymbol{x}^\dagger$, where F and G are a priori chosen proper, convex and l.s.c. functions. In a supervised learning environment, we have training data $\mathcal{S}$ available which comprises a certain number of pairs $(\tilde{\boldsymbol{x}}_t, \boldsymbol{x}_t^\dagger) \in \mathbb{R}^n \times \mathbb{R}^n$. By means of a smooth loss function $L : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, we measure the discrepancy $L(\hat{\boldsymbol{x}}, \boldsymbol{x}^\dagger)$ between a ground truth and the associated reconstruction. Therewith, our goal is to choose $\boldsymbol{K}$ as a solution of the following bilevel optimization problem [1]:

$$\min_{\boldsymbol{K}\in\mathbb{R}^{m\times n}} \sum_{(\tilde{\boldsymbol{x}}_t, \boldsymbol{x}_t^\dagger)\in\mathcal{S}} L(\hat{\boldsymbol{x}}_t, \boldsymbol{x}_t^\dagger) \quad \text{s.t. } \forall t : \hat{\boldsymbol{x}}_t \; \in \; S(\boldsymbol{K}, \tilde{\boldsymbol{x}}_t) \tag{2}$$

II. Plain unrolling scheme

After the change of variables $\boldsymbol{r} \coloneqq \boldsymbol{x} - \tilde{\boldsymbol{x}}$ the Chambolle-Pock iteration [2] for the lower-level problem (1) can be written as

$$\boldsymbol{z}_D^k = \boldsymbol{y}^{k-1} + \sigma\boldsymbol{K}(\tilde{\boldsymbol{x}} + \overline{\boldsymbol{r}}^{k-1}) \qquad \boldsymbol{y}^k = \operatorname{prox}_{\sigma F^*}(\boldsymbol{z}_D^k) \tag{3}$$

$$\boldsymbol{z}_P^k = \boldsymbol{r}^{k-1} - \tau\boldsymbol{K}^\top\boldsymbol{y}^k \qquad \boldsymbol{r}^k = \operatorname{prox}_{\tau G}(\boldsymbol{z}_P^k) \tag{4}$$

$$\overline{\boldsymbol{r}}^k = \boldsymbol{r}^k + \theta(\boldsymbol{r}^k - \boldsymbol{r}^{k-1}) \tag{5}$$

where we use $\boldsymbol{r}^0 = \overline{\boldsymbol{r}}^0 = \boldsymbol{0}$ and $\boldsymbol{y}^0 = \boldsymbol{0}$, and σ, τ and θ are the usual step size and extrapolation parameters of the Chambolle-Pock algorithm. Additionally, we assume that both proximal operators are at least semismooth. Now, the original constraint $\hat{\boldsymbol{x}} \in S(\boldsymbol{K}, \tilde{\boldsymbol{x}})$ is replaced by a relaxed version $\hat{\boldsymbol{x}} = \tilde{\boldsymbol{x}} + \boldsymbol{r}^d(\boldsymbol{K}, \tilde{\boldsymbol{x}})$ which is computed according to (3)–(5) with some fixed $d \in \mathbb{N}$. The relaxed version of the bilevel problem then reads

$$\min_{\boldsymbol{K}\in\mathbb{R}^{m\times n}} \sum_{(\tilde{\boldsymbol{x}}_t, \boldsymbol{x}_t^\dagger)\in\mathcal{S}} L(\tilde{\boldsymbol{x}}_t + \boldsymbol{r}_t^d(\boldsymbol{K}, \tilde{\boldsymbol{x}}_t), \boldsymbol{x}_t^\dagger) \tag{6}$$

and because L is differentiable and the proximal operators are semismooth, the derivative of (6) can be computed explicitly. To that end, it is helpful to interpret the intermediate iterates $\boldsymbol{y}^k$ and $\boldsymbol{r}^k$ as activations of a recurrent neural network $\boldsymbol{r}^d$ with nonlinearities $\operatorname{prox}_{\sigma F^*}$ and $\operatorname{prox}_{\tau G}$ and linear operator $\boldsymbol{K}$. Given that both proximal operators and the respective derivatives can be evaluated element-wise, one can use a backpropagation scheme [3] or an adjoint state method to derive the results presented in the next section.

III. Theoretical results

Lemma 1. *Let $L(\hat{\boldsymbol{x}}, \boldsymbol{x}^\dagger) = \frac{1}{2}\|\hat{\boldsymbol{x}} - \boldsymbol{x}^\dagger\|_2^2$, $\boldsymbol{\delta}_D^{d+1} = \overline{\boldsymbol{\delta}}_D^{d+1} = \boldsymbol{0}$ and $\boldsymbol{\delta}_P^{d+1} = \tilde{\boldsymbol{x}} + \boldsymbol{r}^d(\boldsymbol{K}, \tilde{\boldsymbol{x}}) - \boldsymbol{r}^\dagger$. Moreover, consider the reverse iteration*

$$\boldsymbol{\delta}_P^k = \operatorname{prox}'_{\tau G}(\boldsymbol{z}_P^k) \odot (\boldsymbol{\delta}_P^{k+1} + \sigma\boldsymbol{K}^\top\overline{\boldsymbol{\delta}}_D^{k+1}) \tag{7}$$

$$\boldsymbol{\delta}_D^k = \operatorname{prox}'_{\sigma F^*}(\boldsymbol{z}_D^k) \odot (\boldsymbol{\delta}_D^{k+1} - \tau\boldsymbol{K}\boldsymbol{\delta}_P^k) \tag{8}$$

$$\overline{\boldsymbol{\delta}}_D^k = \boldsymbol{\delta}_D^k + \theta(\boldsymbol{\delta}_D^k - \boldsymbol{\delta}_D^{k+1}) \tag{9}$$

for $k = d, \dots, 1$. Then, the gradient of $L(\tilde{\boldsymbol{x}} + \boldsymbol{r}^d(\boldsymbol{K}, \tilde{\boldsymbol{x}}), \boldsymbol{x}^\dagger)$ with respect to $\boldsymbol{K}$ is given by

$$\sum_{k=1}^{d} \sigma\boldsymbol{\delta}_D^k(\tilde{\boldsymbol{x}} + \overline{\boldsymbol{r}}^{k-1})^\top - \tau\boldsymbol{y}^{k-1}(\boldsymbol{\delta}_P^k)^\top \; . \tag{10}$$

Lemma 1 shows that the sequence generating $\boldsymbol{\delta}_P^k$ and $\boldsymbol{\delta}_D^k$ is almost another instance of Chambolle-Pock, except that the proximal operators are replaced by element-wise multiplication with the respective derivatives. Assuming that these remain constant after a certain number of iterations allows one to show convergence and draw conclusions about limits via the underlying optimization problems.

Theorem 2. *Let $\boldsymbol{K}$, σ and τ satisfy $\sigma\tau\|\boldsymbol{K}\|^2 < 1$ and let $\theta = 1$. Moreover, suppose that there exists a k_0 such that $\operatorname{prox}'_{\tau G}(\boldsymbol{z}_P^k) = \operatorname{prox}'_{\tau G}(\boldsymbol{z}_P^{k_0}) \in \{0,1\}^n$ and $\operatorname{prox}'_{\sigma F^*}(\boldsymbol{z}_D^k) = \operatorname{prox}'_{\sigma F^*}(\boldsymbol{z}_D^{k_0}) \in \{0,1\}^m$ is true for all $k \geq k_0$. Then, it holds that*

$$\lim_{d\to\infty} \boldsymbol{\delta}_P^{k_0} \in \ker(\boldsymbol{K}) \quad \text{and} \quad \lim_{d\to\infty} \boldsymbol{\delta}_D^{k_0} \in \ker(\boldsymbol{K}^\top) \; .$$

IV. Motivation of Ergodic unrolling schemes

Theorem 2 supports that $\boldsymbol{\delta}_P^k$ and $\boldsymbol{\delta}_D^k$ can vanish, especially in case d is large enough to ensure convergence of (3)–(5). However, this behavior may be undesirable, e.g. if we want to learn an operator $\boldsymbol{K}$ that grants fast progress in early iterations of the forward scheme, or because the minimization of (6) can get stuck in bad local minima. This motivates us to adapt (6), this time using weighted ergodic averages $\boldsymbol{e}^d = \sum_{k=1}^d \alpha_k \boldsymbol{r}^k$ instead of only the final iterate. The effect on (7)–(9) can be considered a summation of several backpropagation sequences starting at each iterate $\boldsymbol{r}^k$ with $k \leq d$. This modification theoretically prevents from the behavior described in Theorem 2, and our experiments show that it can also pay off in terms of substantially lower loss values on the training data.

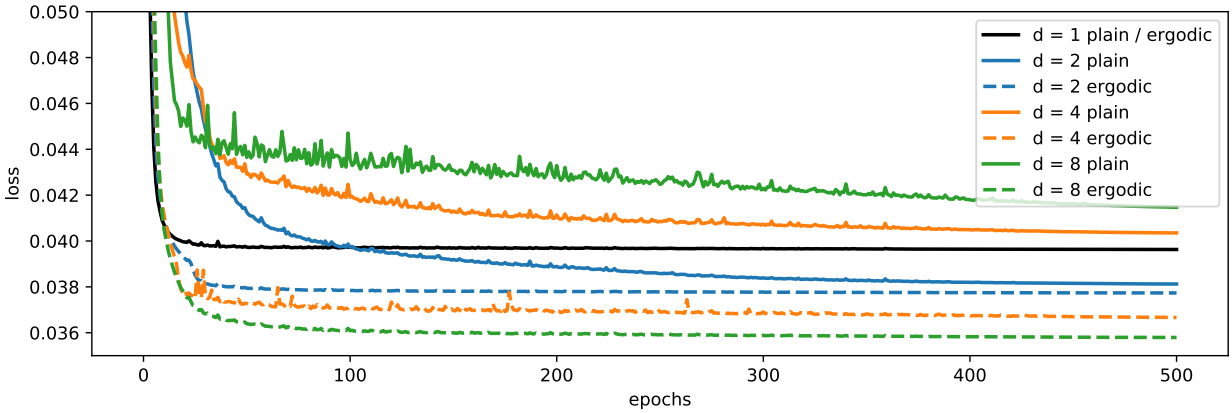

Fig. 1. We take up the experimental setting introduced in [4], [5], [6] in the context of speech dequantization, where F ist the ℓ_1-norm, G is an indicator function encoding the constraint $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_\infty \leq \eta$, and the training data consists of 66.628 ground truth speech signal snippets form the IEEE corpus [7] and associated quantized versions. Both proximal operators are then projections onto sup norm balls satisfying the assumptions for Lemma 1 and Theorem 2. Throughout, we use stochastic gradient descent with decreasing step sizes to minimize the loss (6). For ergodic averaging we use weights $\alpha_k = \mathcal{O}(1/k)$. In case $d = 1$, we set $\alpha_1 = 1$ such that the plain and the ergodic setting produce identical results. In the remaining cases with $d \in \{2, 4, 8\}$ it can be observed that an increasing number of unrolled iterations leads to increasing loss values in the plain setting, whereas ergodic averaging leads to decreasing loss values induced by the learned analysis operators.

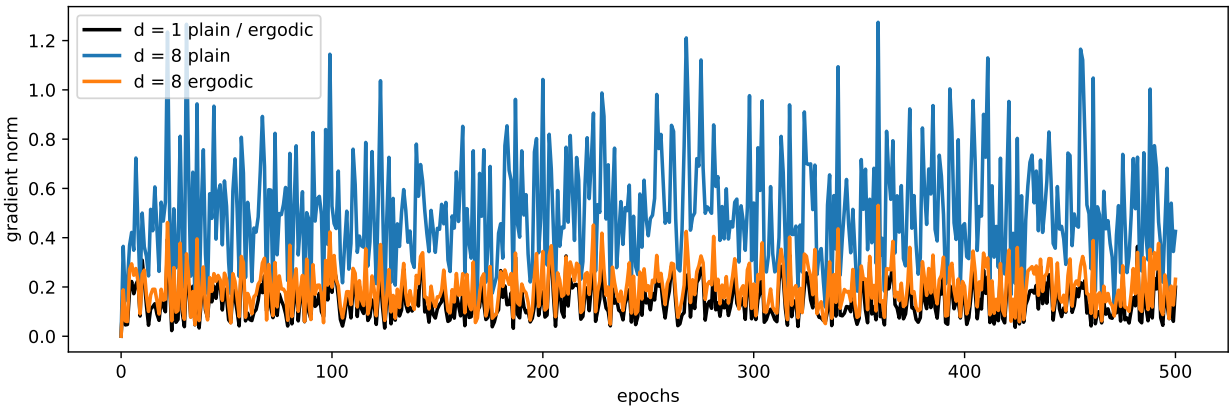

Fig. 2. In the same experimental setting as in Figure 1, it can be observed that the norms of the gradients (10) seem to be less dependent on the number of unrolled iterations in case one uses ergodic averaging. This behavior supports a more stable behavior of the training procedure and more robustness with respect to the utilized step sizes for stochastic gradient descent.

References

[1] B. Colson, P. Marcotte and G. Savard, *An overview of bilevel optimization*, Annals of Operations Research 153(1), pp. 235–256, 2007.

[2] A. Chambolle and T. Pock, *A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging*, Journal of Mathematical Imaging and Vision 40(1), pp. 120–145, 2011.

[3] D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Learning internal representations by error propagation*, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[4] C. Brauer, T. Gerkmann and D. Lorenz, *Sparse reconstruction of quantized speech signals*, In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5940–5944, 2016.

[5] C. Brauer and D. Lorenz, *Primal-dual residual networks*, arXiv preprint arXiv:1806.05823, 2018.

[6] C. Brauer, Z. Zhao, D. Lorenz and T. Fingscheidt, *Learning to dequantize speech signals by primal-dual networks: An approach for acoustic sensor networks*, To appear in: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[7] P. C. Loizou, *Speech Enhancement - Theory and Practice*, CRC Press, Taylor & Francis Group, 2007.

[8] J. Adler and O. Öktem, *Learned Primal-dual Reconstruction*, IEEE Transactions on Medical Imaging 37(6), pp. 1322-1332, 2018.

[9] A. Chambolle and T. Pock, *On the ergodic convergence rates of a first-order primal–dual algorithm*, Mathematical Programming 150(1-2), pp. 253–287, 2016.

[10] Y. Chen, T. Pock and H. Bischof, *Learning ℓ_1-based analysis and synthesis sparsity priors using bi-level optimization*, arXiv preprint arXiv:1401.4105, 2014.

[11] K. Gregor and Y. LeCun, *Learning Fast Approximations of Sparse Coding*, In: Proceedings of the 27th International Conference on Machine Learning (ICML), pp. 399–406, 2010.

[12] K. Kunisch and T. Pock, *A bilevel optimization approach for parameter learning in variational models*, SIAM Journal on Imaging Sciences 6(2), pp. 938–983, 2013.

[13] P. Ochs, R. Ranftl, R. Brox and T. Pock, *Bilevel Optimization with Nonsmooth Lower Level Problems*, In: International Conference on Scale Space and Variational Methods in Computer Vision (SSVM), pp. 654–665, 2015.

[14] P. Ochs, R. Ranftl, R. Brox and T. Pock, *Techniques for Gradient-Based Bilevel Optimization with Non-smooth Lower Level Problems*, Journal of Mathematical Imaging and Vision 56(2), pp. 175–194, 2016.

[15] G. Riegler, D. Ferstl, M. Rüther and H. Bischof, *A Deep Primal-Dual Network for Guided Depth Super-Resolution*, In: Proceedings of the British Machine Vision Conference (BMVC), pp. 7.1–7.14, 2016.

[16] G. Riegler, M. Rüther and H. Bischof, *ATGV-Net: Accurate Depth Super-Resolution*, In: European Conference on Computer Vision (ECCV), pp. 268–284, 2016.

[17] S. Wang, S. Fidler and R. Urtasun, *Proximal Deep Structured Models*, In: Advances in Neural Information Processing Systems (NIPS), pp. 865–873, 2016.