Technische Universität Braunschweig

SPONSORED BY THE
Federal Ministry of Education and Research

DLR
Institut für Systemleichtbau

EUSIPCO2023

Asymptotic analysis and truncated backpropagation for the unrolled primal-dual algorithm

Christoph Brauer and Dirk Lorenz, September 8, 2023

Introduction

Task

Recover ground truth $\boldsymbol{y} \in \mathbb{R}^n$
from noisy observation $\boldsymbol{x} \in \mathbb{R}^n$

use $\hat{\boldsymbol{K}}$

Bilevel Problem

$$\begin{aligned} \widehat{\boldsymbol{K}} \in \operatorname*{argmin}_{\boldsymbol{K}} \quad & \sum_{i=1}^{m} \ell(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i) \\ \text{s.t.} \quad & \forall i : \hat{\boldsymbol{y}}_i \in S(\boldsymbol{K}, \boldsymbol{x}_i) \end{aligned}$$

via

via

Convex Problem

$$\hat{\boldsymbol{y}} \in S(\boldsymbol{K}, \boldsymbol{x}) := \operatorname*{argmin}_{\boldsymbol{y}} F(\boldsymbol{K}\boldsymbol{y}) + G(\boldsymbol{y} - \boldsymbol{x})$$

learn $\boldsymbol{K}$

from

Training Data

$$\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_m, \boldsymbol{y}_m)\}$$

Introduction

Bilevel Problem

$$\begin{aligned} \widehat{\boldsymbol{K}} \in \underset{\boldsymbol{K}}{\operatorname{argmin}} \quad & \sum_{i=1}^{m} \ell(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i) \\ \text{s.t.} \quad & \forall i : \hat{\boldsymbol{y}}_i \in S(\boldsymbol{K}, \boldsymbol{x}_i) \end{aligned}$$

Convex Optimization Algorithm

$$A^1(\boldsymbol{K}, \boldsymbol{x}), \ldots, A^L(\boldsymbol{K}, \boldsymbol{x})$$

Approximate Bilevel Problem

$$\widehat{\boldsymbol{K}} \in \underset{\boldsymbol{K}}{\operatorname{argmin}} \sum_{i=1}^{m} \ell(\boldsymbol{y}_i, A^L(\boldsymbol{K}, \boldsymbol{x}_i))$$

Introduction

Approximate Bilevel Problem

$$\widehat{\boldsymbol{K}} \in \operatorname*{argmin}_{\boldsymbol{K}} \sum_{i=1}^{m} \ell(\boldsymbol{y}_i, A^L(\boldsymbol{K}, \boldsymbol{x}_i))$$

single

example

Approximate Gradient

$$\nabla_{\boldsymbol{K}}\, \ell(\boldsymbol{y}, A^L(\boldsymbol{K}, \boldsymbol{x})) \xrightarrow[\text{for } L \to \infty]{\text{behavior}} \text{?}$$

Outline

1. The algorithm A
2. Gradient backpropagation
3. Gradient of $\ell(\boldsymbol{y}, A^L(\boldsymbol{K}, \boldsymbol{x}))$ w.r.t. parameters $\boldsymbol{K}$
4. Limit of parameter gradient for $L \to \infty$
5. Speech dequantization
6. Truncated Backpropagation
7. Interpretability

The algorithm A

Algorithm 1 Chambolle-Pock

Choose $\sigma, \tau > 0$ and $\theta \in [0, 1]$
Initialize $\boldsymbol{y}^0 = \bar{\boldsymbol{y}}^0 = \boldsymbol{x}$ and $\boldsymbol{\psi}^0 = 0$
for $l = 0, \ldots, L-1$ **do**

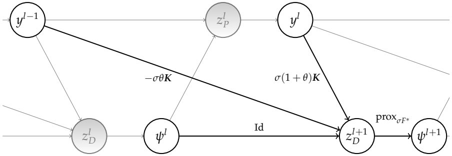
$$\left.\begin{aligned} \boldsymbol{z}_D^{l+1} &= \boldsymbol{\psi}^l + \sigma \boldsymbol{K} \bar{\boldsymbol{y}}^l \\ \boldsymbol{\psi}^{l+1} &= \operatorname{prox}_{\sigma F^*}(\boldsymbol{z}_D^{l+1}) \end{aligned}\right\} \quad \text{(Dual update)}$$

$$\left.\begin{aligned} \boldsymbol{z}_P^{l+1} &= \boldsymbol{y}^l - \tau \boldsymbol{K}^\top \boldsymbol{\psi}^{l+1} \\ \boldsymbol{y}^{l+1} &= \operatorname{prox}_{\tau G}(\boldsymbol{z}_P^{l+1} - \boldsymbol{x}) + \boldsymbol{x} \end{aligned}\right\} \quad \text{(Prim. update)}$$

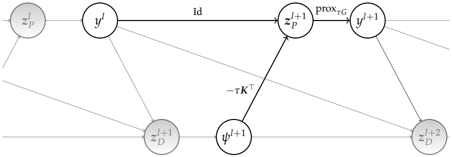
$$\bar{\boldsymbol{y}}^{l+1} = \boldsymbol{y}^{l+1} + \theta(\boldsymbol{y}^{l+1} - \boldsymbol{y}^l) \quad \text{(Extrapolation)}$$

end for

Dual update



Primal update



$\rightarrow$ Apply standard backprop as for FCNNs

Gradient backpropagation

Algorithm 2 Backpropagated gradients

Choose σ, τ and θ as in Algorithm 1
Adopt $\boldsymbol{y}^L, \boldsymbol{z}_P^1, \dots, \boldsymbol{z}_P^L, \boldsymbol{z}_D^1, \dots, \boldsymbol{z}_D^L$ from Algorithm 1
Initialize $\delta_P^{L+1} = \nabla_{\hat{\boldsymbol{y}}}\, \ell(\boldsymbol{y}, \boldsymbol{y}^L)$
Initialize $\delta_D^{L+1} = \overline{\delta}_D^{L+1} = 0$
for $l = L, \dots, 1$ **do**

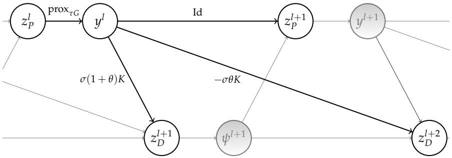
$$\delta_P^l = \mathcal{J}_{\mathrm{prox}_{\tau G}}(\boldsymbol{z}_P^l - \boldsymbol{x})^\top (\delta_P^{l+1} + \sigma \boldsymbol{K}^\top \overline{\delta}_D^{l+1}) \tag{1}$$

$$\delta_D^l = \mathcal{J}_{\mathrm{prox}_{\sigma F^*}}(\boldsymbol{z}_D^l)^\top (\delta_D^{l+1} - \tau \boldsymbol{K} \delta_P^l) \tag{2}$$

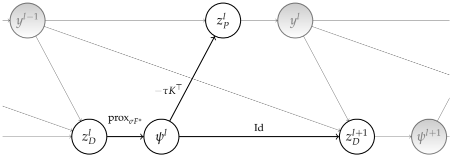
$$\overline{\delta}_D^l = \delta_D^l + \theta(\delta_D^l - \delta_D^{l+1}) \tag{3}$$

end for

Primal gradient $\delta_P^l := \nabla_{\boldsymbol{z}_P^l}\, \ell(\boldsymbol{y}, \boldsymbol{y}^L)$



Dual gradient $\delta_D^l := \nabla_{\boldsymbol{z}_D^l}\, \ell(\boldsymbol{y}, \boldsymbol{y}^L)$



$\rightarrow$ Basically chain rule

Parameter gradient

Lemma

The gradients

$$\delta_P^l := \nabla_{\mathbf{z}_P^l} \ell(\mathbf{y}, \mathbf{y}^L) \quad \textit{and}$$
$$\delta_D^l := \nabla_{\mathbf{z}_D^l} \ell(\mathbf{y}, \mathbf{y}^L)$$

can be computed recursively as outlined in Algorithm 2. Moreover, it holds that

$$\nabla_{\boldsymbol{K}} \ell(\mathbf{y}, \mathbf{y}^L) = \sum_{l=1}^{L} \sigma \delta_D^l (\bar{\mathbf{y}}^{l-1})^\top - \tau \boldsymbol{\psi}^l (\delta_P^l)^\top .$$

Proof.

$\boldsymbol{K}$ affects the objective exactly through $\mathbf{z}_P^1, \ldots, \mathbf{z}_P^L$ and $\mathbf{z}_D^1, \ldots, \mathbf{z}_D^L$. $\rightarrow$ Chain rule. $\square$

Limit of parameter gradient

Theorem

Suppose that there exist constants $c \geq 0$ and $0 \leq \kappa < 1$ such that

$$\|\delta_D^{l,L}\| \leq c\kappa^{L-l} \quad \text{and} \quad \|\delta_P^{l,L}\| \leq c\kappa^{L-l}$$

hold for arbitrary L and $l \in \{1, \ldots, L\}$. Then, the limits

$$\Delta_P := \lim_{L\to\infty} \sum_{l=1}^{L} \delta_P^{l,L} \quad \text{and} \quad \Delta_D := \lim_{L\to\infty} \sum_{l=1}^{L} \delta_D^{l,L}$$

exist and are finite. Moreover, it holds that

$$\lim_{L\to\infty} \nabla_{\boldsymbol{\kappa}}\, \ell(\boldsymbol{y}, \boldsymbol{y}^L) = \sigma\Delta_D(\boldsymbol{y}^*)^\top - \tau\boldsymbol{\psi}^*(\Delta_P)^\top .$$

Limit of parameter gradient

Proof.

1. Show that Δ_P and Δ_D exist and are finite

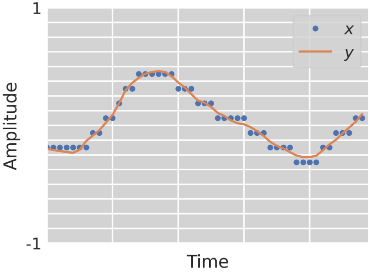
$$\lim_{L\to\infty}\sum_{l=1}^{L}\|\delta_P^{l,L}\| \leq \lim_{L\to\infty}\sum_{l=1}^{L} c\kappa^{L-l} = \lim_{L\to\infty} c\,\frac{1-\kappa^L}{1-\kappa} = \frac{c}{1-\kappa} < \infty$$

2. Rearrange gradient formula

$$\begin{aligned}
\nabla_K \ell(y, y^L) &= \sum_{l=1}^{L} \sigma\delta_D^{l,L}(\bar{y}^{l-1})^\top - \tau\psi^l(\delta_P^{l,L})^\top \\
&= \sum_{l=1}^{L} \sigma\delta_D^{l,L}(y^*)^\top - \tau\psi^*(\delta_P^{l,L})^\top - \sigma\delta_D^{l,L}(y^* - \bar{y}^{l-1})^\top + \tau(\psi^* - \psi^l)(\delta_P^l)^\top \\
&= \underbrace{\sigma\Big(\sum_{l=1}^{L}\delta_D^{l,L}\Big)(y^*)^\top - \tau\psi^*\Big(\sum_{l=1}^{L}\delta_P^{l,L}\Big)^\top}_{\xrightarrow{L\to\infty}\sigma\Delta_D(y^*)^\top - \tau\psi^*(\Delta_P)^\top} - \underbrace{\sigma\sum_{l=1}^{L}\delta_D^{l,L}(y^* - \bar{y}^{l-1})^\top}_{\xrightarrow{L\to\infty} 0} + \underbrace{\tau\sum_{l=1}^{L}(\psi^* - \psi^l)(\delta_P^{l,L})^\top}_{\xrightarrow{L\to\infty} 0}
\end{aligned}$$

$\square$

Speech dequantization



$$\hat{\boldsymbol{y}} \in \operatorname*{argmin}_{\boldsymbol{y}} \underbrace{\|\boldsymbol{K}\boldsymbol{y}\|_1}_{F(\boldsymbol{K}\boldsymbol{y})} \quad \underbrace{\text{s.t. } \|\boldsymbol{y}-\boldsymbol{x}\|_\infty \leq \frac{\eta}{2}}_{G(\boldsymbol{y}-\boldsymbol{x}) = I_{\|\cdot\|_\infty \leq \frac{\eta}{2}}(\boldsymbol{y}-\boldsymbol{x})}$$

$$\boldsymbol{K} \in \mathbb{R}^{k \times n}$$

$$F^* = I_{\|\cdot\|_\infty \leq 1}$$

$$\operatorname{prox}_{\sigma F^*}(\boldsymbol{z}_D^l) = \min\left\{1, \max\left\{-1, \boldsymbol{z}_D^l\right\}\right\}$$

$$\operatorname{prox}_{\tau G}(\boldsymbol{z}_P^l - \boldsymbol{x}) = \min\left\{\frac{\eta}{2}, \max\left\{-\frac{\eta}{2}, \boldsymbol{z}_P^l - \boldsymbol{x}\right\}\right\}$$

$$\delta_P^l = \overbrace{\operatorname{prox}'_{\tau G}(\boldsymbol{z}_P^l - \boldsymbol{x})}^{\in\{0,1\}^n} \odot (\delta_P^{l+1} + \sigma \boldsymbol{K}^\top \bar{\delta}_D^{l+1})$$

$$\delta_D^l = \underbrace{\operatorname{prox}_{\sigma F^*}(\boldsymbol{z}_D^l)}_{\in\{0,1\}^k} \odot (\delta_D^{l+1} - \tau \boldsymbol{K} \delta_P^l)$$

Speech dequantization

Theorem

Suppose that there exists an $l_0 \in \mathbb{N}$ *such that* $\operatorname{prox}'_{\tau G}(\mathbf{z}_P^l - \boldsymbol{x}) = \operatorname{prox}'_{\tau G}(\mathbf{z}_P^{l_0} - \boldsymbol{x})$ *and* $\operatorname{prox}'_{\sigma F^*}(\mathbf{z}_D^l) = \operatorname{prox}'_{\sigma F^*}(\mathbf{z}_D^{l_0})$ *hold for all* $l \geq l_0$.

Then, it holds for all $l \geq l_0$ *that* $\lim_{L\to\infty} \delta_P^l \in \ker(\boldsymbol{K})$ *and* $\lim_{L\to\infty} \delta_D^l \in \ker(\boldsymbol{K}^\top)$.

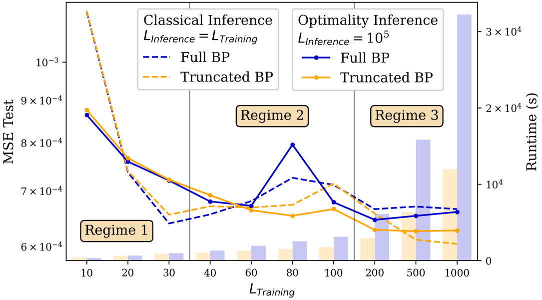
Proof.

Set $\tilde{\boldsymbol{K}} := \operatorname{prox}'_{\sigma F^*}(\mathbf{z}_D^{l_0}) \odot \boldsymbol{K} \odot \operatorname{prox}'_{\tau G}(\mathbf{z}_P^{l_0} - \boldsymbol{x})$ and reverse engineer the optimization problem behind the Chambolle-Pock iteration to see that $\lim_{L\to\infty} \delta_D^l \in \operatorname{argmin}_{\delta_D} \ \text{const.} \quad \text{s.t.} \quad \tilde{\boldsymbol{K}}^\top \delta_D = \mathbf{0}$ and analogously for the other case.

$$\delta_P^l = \delta_P^{l+1} + \sigma \tilde{\boldsymbol{K}}^\top \bar{\delta}_D^{l+1}$$
$$\delta_D^l = \delta_D^{l+1} - \tau \tilde{\boldsymbol{K}} \delta_P^l$$

$\square$

Truncated Backprop



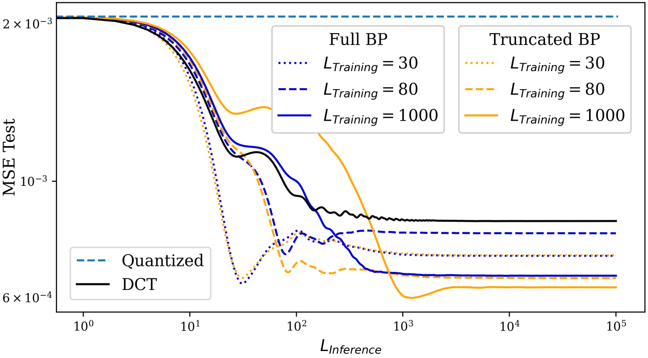
Regime 1. Full and truncated BP similar

Regime 2. No significant performance increase

Regime 3. Truncated BP outperforms full BP

- Full BP
  - Full backward pass according to Lemma
  - Requires $\sim 2L \times$ runtime of single iteration
  - Requires storage of $2L(k+n)$ variables
- Truncated BP
  - Use only $\boldsymbol{y}^L, \boldsymbol{\psi}^L, \boldsymbol{z}_P^L, \boldsymbol{z}_D^L$ and finite number of b backpropagated gradients to approximate $\lim_{L\to\infty} \nabla_{\boldsymbol{K}} \ell(\boldsymbol{y}, \boldsymbol{y}^L) = \sigma \Delta_D (\boldsymbol{y}^*)^\top - \tau \boldsymbol{\psi}^* (\Delta_P)^\top$
  - Requires $\sim L + b \times$ runtime of single iteration
  - Requires storage of $2(k+n)$ variables

Interpretability



- Bump at 30 iterations illustrates gap between classical and optimality inference in Regime 1
- Classical unrolling is prone to overfitting to the number of unrolled iterations
- The best performing model in terms of the error here also features the slowest convergence speed

Summary

- Asymptotic Analysis. Limit of parameter gradient depends only on optimal solutions
- Truncated backpropagation. Outperforms usage of full gradients
- Interpretability. Limited when too few unrolled iterations during training
- Future work. Different algorithms, combine high interpretability and convergence speed

`https://github.com/chrbraue/primal_dual_networks`

Thank you!