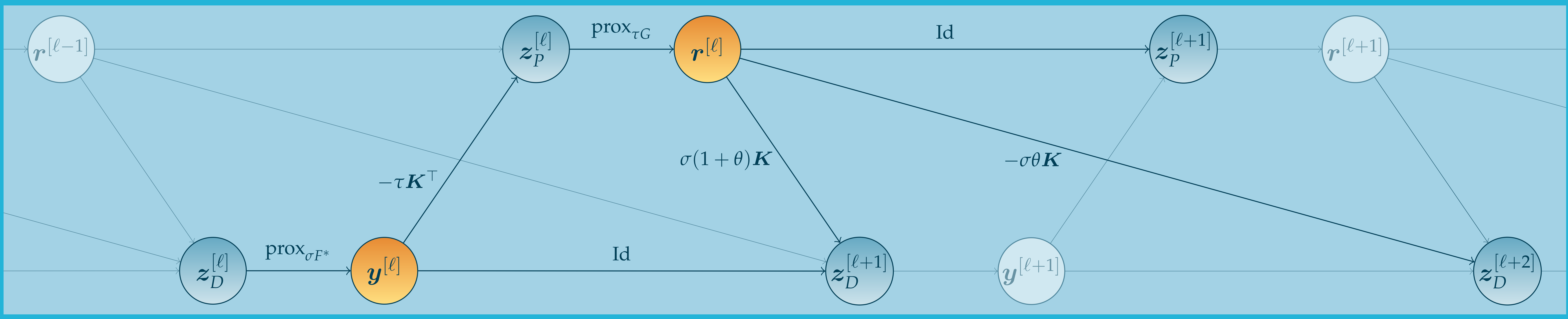
Technische Universität Braunschweig

Asymptotic analysis of unrolling approaches for bi-level optimization

Christoph Brauer and Dirk Lorenz

Technische Universität Braunschweig | Institute of Analysis and Algebra



Introduction

We want to find an analysis operator K that solves the following bilevel optimization problem:

$$\begin{aligned}&\underset{K\in\mathbb{R}^{m\times n}}{\operatorname{argmin}} \quad \sum_i \ell(\hat{x}_i, x_i^\dagger)\\ &\quad\text{s.t.}\quad \forall i: \hat{x}_i \in \underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\ F(Kx) + G(x-\tilde{x}_i)\end{aligned}$$

To that end, we substitute $r := x - \tilde{x}$ and apply Chambolle-Pock to the lower-level problem $\tilde{x} + \operatorname{argmin}_{r\in\mathbb{R}^n} F(Kr + K\tilde{x}) + G(r)$:

$$\begin{aligned} z_D^{[\ell+1]} &:= y^{[\ell]} + \sigma K(\tilde{x} + \bar{r}^{[\ell]}) & y^{[\ell+1]} &:= \mathrm{prox}_{\sigma F^*}(z_D^{[\ell+1]})\\ z_P^{[\ell+1]} &:= r^{[\ell]} - \tau K^\top y^{[\ell+1]} & r^{[\ell+1]} &:= \mathrm{prox}_{\tau G}(z_P^{[\ell+1]})\\ & & \bar{r}^{[\ell+1]} &:= r^{[\ell+1]} + \theta(r^{[\ell+1]} - r^{[\ell]})\end{aligned}$$

Finally, we fix $L \in \mathbb{N}$ and replace the constraints in the bilevel problem by the approximation $\hat{x} = \tilde{x} + A(K,\tilde{x}) := \tilde{x} + r^{[L]}$. Hence, the bilevel problem can be rewritten unconstrained. Given that the proximal operators are sufficiently smooth, we can also take derivatives with respect to K and thus apply gradient descent.

One drawback of this approach is that the number L of unrolled iterations often needs to be high in order to achieve a good approximation of the argmin operator. This can be a problem in terms of required computational resources.

Exact gradient computation

$A(K,\tilde{x})$ is a specific recurrent neural network. Using backprop we show that $\delta_P^{[\ell]} := \nabla_{z_P^{[\ell]}} \ell(\tilde{x} + r^{[L]}, x^\dagger)$ and $\delta_D^{[\ell]} := \nabla_{z_D^{[\ell]}} \ell(\tilde{x} + r^{[L]}, x^\dagger)$ can be computed recursively for $\ell = L, \dots, 1$:

$$\begin{aligned}\delta_P^{[\ell]} &= \mathrm{prox}'_{\tau G}(z_P^{[\ell]}) \odot (\delta_P^{[\ell+1]} + \sigma K^\top \bar{\delta}_D^{[\ell+1]})\\ \delta_D^{[\ell]} &= \mathrm{prox}'_{\sigma F^*}(z_D^{[\ell]}) \odot (\delta_D^{[\ell+1]} - \tau K \delta_P^{[\ell]})\\ \bar{\delta}_D^{[\ell]} &= \delta_D^{[\ell]} + \theta(\delta_D^{[\ell]} - \delta_D^{[\ell+1]})\end{aligned}$$

The gradient with respect to the parameters is then:

$$\nabla_K \ell(\tilde{x} + r^{[L]}, x^\dagger) = \sum_{\ell=1}^{L} \sigma \delta_D^{[\ell]} (\tilde{x} + \bar{r}^{[\ell-1]})^\top - \tau y^{[\ell]} \delta_P^{[\ell]\top}$$

Asymptotical gradient

We assume that

$$\Delta_P := \lim_{L\to\infty} \sum_{\ell=1}^{L} \delta_P^{[\ell]} < \infty \quad \text{and} \quad \Delta_D := \lim_{L\to\infty} \sum_{\ell=1}^{L} \delta_D^{[\ell]} < \infty$$

and define

$$x^* := \tilde{x} + \lim_{\ell\to\infty} r^{[\ell]} \quad \text{and} \quad y^* := \lim_{\ell\to\infty} y^{[\ell]}\,.$$

If

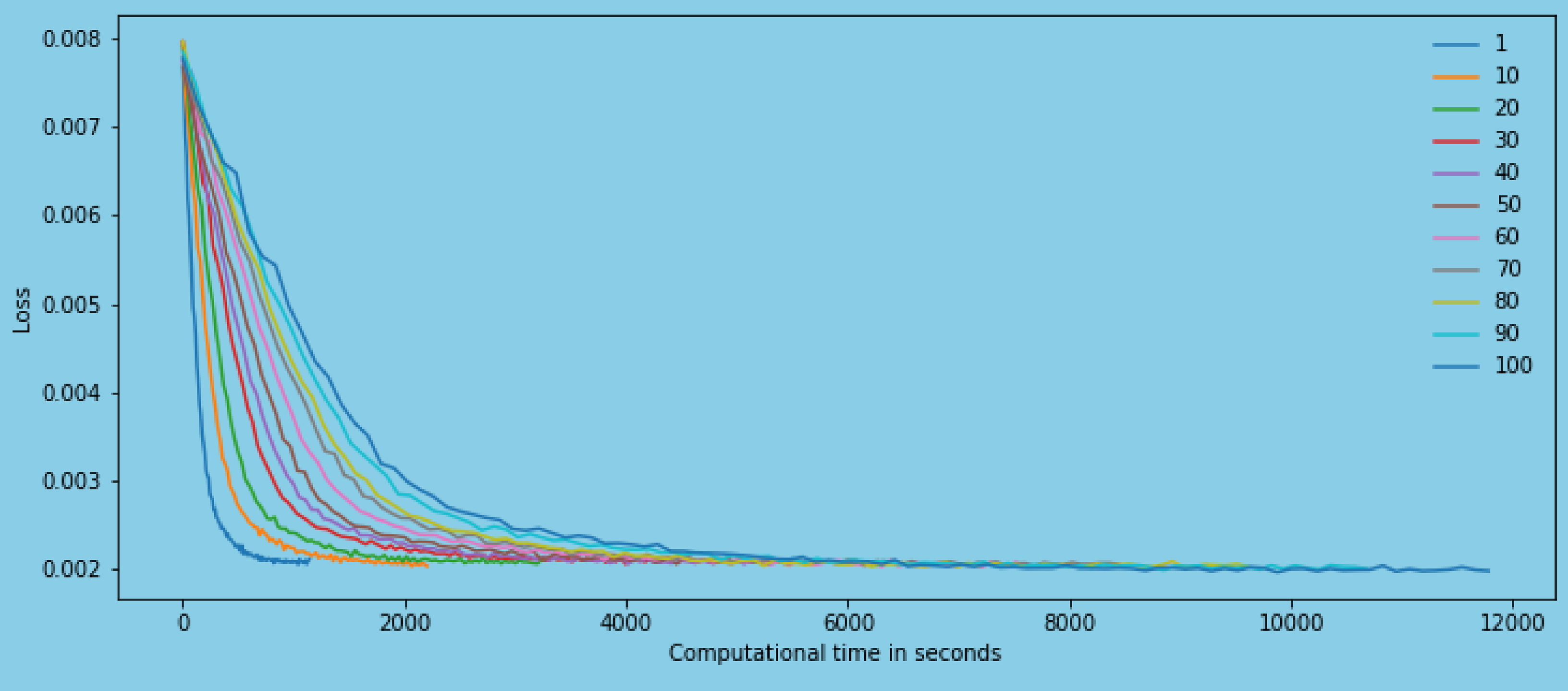
$$\lim_{L\to\infty} \sum_{\ell=1}^{L} |\delta_P^{[\ell]}| < \infty \quad \text{and} \quad \lim_{L\to\infty} \sum_{\ell=1}^{L} |\delta_D^{[\ell]}| < \infty\,,$$

then it holds that

$$\lim_{L\to\infty} \nabla_K \ell(\tilde{x} + A^L(K,\tilde{x}), x^\dagger) = \sigma \Delta_D x^{*\top} - \tau y^* \Delta_P^\top\,.$$

Early experiments

In early experiments, we investigate the effect of approximating Δ_D and Δ_P by backpropagating only a few steps, i.e. by using only $\delta_D^{[L]}, \dots, \delta_D^{[L-k]}$ and $\delta_P^{[L]}, \dots, \delta_P^{[L-k]}$ for some small k. Moreover, we approximate the primal and dual optimal solutions x^* and y^* using Chambolle-Pock with a number of $L = 100$ iterations.



The legend refers to the number k of backpropagation steps. Our results show that backpropagating for only one step is enough obtain the same overall loss as when backpropagation through the entire unrolled algorithm, while saving a substantial amount of computational resources and time.