Technische Universität Braunschweig

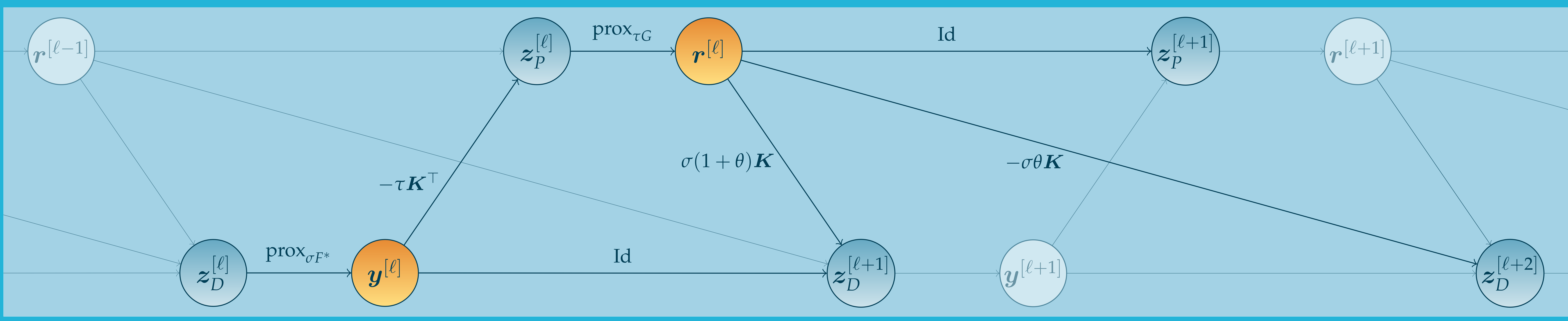
SPARS

Ergodic bilevel optimization

Christoph Brauer und Dirk Lorenz
Technische Universität Braunschweig | Institute for Analysis and Algebra
{ch.brauer, d.lorenz}@tu-braunschweig.de

Takeaway

Learning analysis operators through unrolled Chambolle-Pock iterations is prone to vanishing gradients. Unrolling ergodic averages instead can help to mitigate the problem.



Introduction

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

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

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. Our theoretical results indicate that this approach is prone to vanishing gradients. We propose to unroll ergodic averages

$$e^{[L]} := \sum_{\ell=1}^{L} \alpha_\ell r^{[\ell]}$$

and replace $\hat{x} = \tilde{x} + e^{[L]}$ instead.

Results

$A(K, \tilde{x})$ 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 $\ell = L, \ldots, 1$:

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

The gradient with respect to the parameters is then:

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

Under appropriate conditions there exists an ℓ_0 such that:

$$\lim_{L\to\infty} \delta_P^{[\ell_0]} \in \ker(K) \quad \text{and} \quad \lim_{L\to\infty} \delta_D^{[\ell_0]} \in \ker(K^\top)$$

Unrolling ergodic averages yields lower losses when L is increased: