Adaptive Three Operator Splitting

Fabian Pedregosa[†,‡], Gauthier Gidel[*]

[†]UC Berkeley [‡]ETH Zurich [*]Mila and DIRO, Université de Montréal

Berkeley ETH zürich Mila Université de Montréal

Problem Setting

Goal: solve optimization problems of the form

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}}\; f(\boldsymbol{x}) + g(\boldsymbol{x}) + h(\boldsymbol{x})\,, \qquad \text{(OPT)}$$

with access to ∇f, $\mathbf{prox}_{\gamma g} \stackrel{\text{def}}{=} \arg\min_{\boldsymbol{x}} g(\boldsymbol{x}) + \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{z}\|^2$, $\mathbf{prox}_{\gamma h}$.

- Optimization problems with smooth + non-smooth objectives are ubiquitous in machine learning.
- Many complex penalties can be written as sum of proximable: overlapping group lasso, ℓ_1 trend filtering, isotonic constraints, total variation, intersection of constraints, etc.

Three Operator Splitting

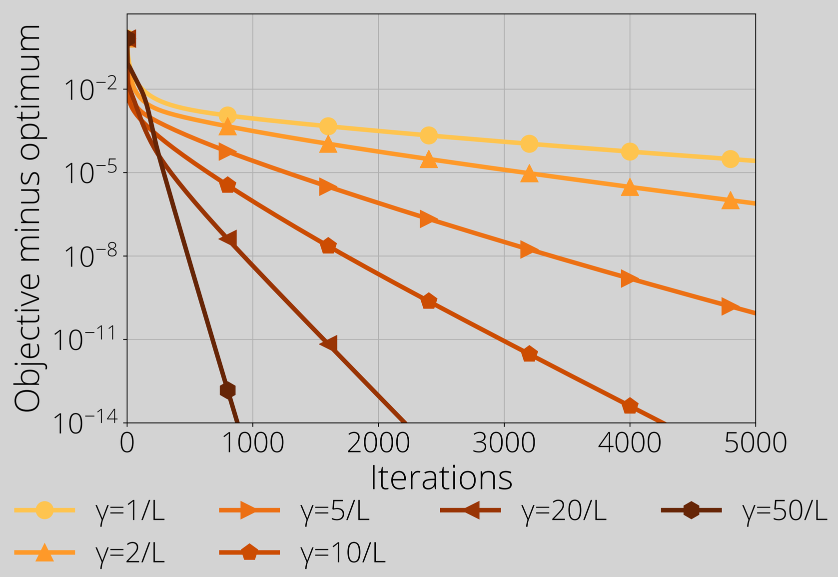
The **Three Operator Splitting** (TOS) is a recently proposed method to solve (OPT)(Davis and Yin 2017). Iterates on d-dim vector $\boldsymbol{y}$:

$$\boldsymbol{z} = \mathbf{prox}_{\gamma h}(\boldsymbol{y})\,,\;\; \boldsymbol{x} = \mathbf{prox}_{\gamma g}(2\boldsymbol{y} - \boldsymbol{z} - \gamma\nabla f(\boldsymbol{z}))\,,\;\; \boldsymbol{y}^+ = \boldsymbol{y} - \boldsymbol{z} + \boldsymbol{x}$$

- Can be used to solve problems with an arbitrary number of proximal terms

$$\min_{\boldsymbol{x}} \varphi(\boldsymbol{x}) + \textstyle\sum_{j=1}^k h_j(\boldsymbol{x}) \equiv \min_{\boldsymbol{X}} \underbrace{\varphi(\overline{\boldsymbol{X}})}_{=f(\boldsymbol{X})} + \underbrace{\textstyle\sum_{j=1}^k h_j(\boldsymbol{X}_j)}_{=h(\boldsymbol{X})} + \underbrace{\imath\{\boldsymbol{X}_1 = \cdots = \boldsymbol{X}_k\}}_{=g(\boldsymbol{X})}\,.$$

- Convergence has been proven for step-sizes $\gamma < 2/L$, with L = Lipschitz constant of ∇f.
- Often best step-size is orders of magnitude larger than theoretical one.
- Can we design a variant that computes step-size based exclusively on local information of the objective?



Contributions

- Three operator splitting with step-size adaptive to local information.
- Step size grows to largest admissible step-size, no hyperparameter to tune.
- Same theoretical guarantees than fixed step-size variant.
- New analysis of TOS based on saddle-point suboptimality: tighter rates and simpler proof.

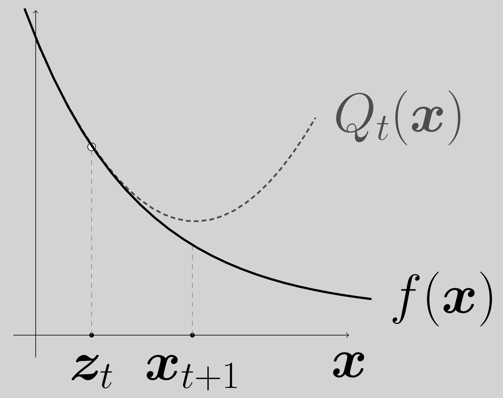
Adaptive Three Operator Splitting

Key idea for adaptive step-size strategy:

- Construct surrogate quadratic of f at $\boldsymbol{z}_t$:

$$Q_t(\boldsymbol{x},\gamma) \stackrel{\text{def}}{=} f(\boldsymbol{z}_t) + \langle\nabla f(\boldsymbol{z}_t), \boldsymbol{x}-\boldsymbol{z}\rangle + \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{z}\|^2\,.$$

- Choose γ_t such that $Q_t(\boldsymbol{x}_{t+1},\gamma_t)$ is upper bound of $f(\boldsymbol{x}_{t+1})$.



Algorithm

1. Start with optimistic step-size γ_t and decrease it until:

$$f(\boldsymbol{x}_{t+1}) \le f(\boldsymbol{z}_t) + \langle\nabla f(\boldsymbol{z}_t), \boldsymbol{x}_{t+1} - \boldsymbol{z}_t\rangle + \frac{1}{2\gamma_t}\|\boldsymbol{x}_{t+1} - \boldsymbol{z}_t\|^2$$

with $\boldsymbol{x}_{t+1} = \mathbf{prox}_{\gamma_t g}(\boldsymbol{z}_t - \gamma_t(\nabla f(\boldsymbol{z}_t) + \boldsymbol{u}_t))$

2. Run the following updates selected step-size:

$$\boldsymbol{z}_{t+1} = \mathbf{prox}_{\gamma_t h}(\boldsymbol{x}_{t+1} + \gamma_t \boldsymbol{u}_t)\,,\quad \boldsymbol{u}_{t+1} = \boldsymbol{u}_{t+1} + (\boldsymbol{x}_{t+1} - \boldsymbol{z}_{t+1})$$

Analysis Framework

⚠ Iterates are not always feasible, objective can be $+\infty$. ⚠

A better gap function emerges from the saddle-point formulation

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) + g(\boldsymbol{x}) + h(\boldsymbol{x}) = \min_{\boldsymbol{x}\in\mathbb{R}^d}\max_{\boldsymbol{u}\in\mathbb{R}^d} \underbrace{f(\boldsymbol{x}) + g(\boldsymbol{x}) + \langle\boldsymbol{x},\boldsymbol{u}\rangle - h^*(\boldsymbol{u})}_{\stackrel{\text{def}}{=}\mathcal{L}(\boldsymbol{x},\boldsymbol{u})}$$

- $\mathcal{L}(\boldsymbol{x}_t,\boldsymbol{u}_t) < \infty$
- $\min_{\boldsymbol{u}} \mathcal{L}(\boldsymbol{x}_t,\boldsymbol{u}) = f(\boldsymbol{x}_t) + g(\boldsymbol{x}_t) + h(\boldsymbol{x}_t)$
- $\mathcal{L}(\boldsymbol{x}^\star,\boldsymbol{u}) - \mathcal{L}(\boldsymbol{x},\boldsymbol{u}^\star) \le 0$ for all $(\boldsymbol{x},\boldsymbol{u})$ $\iff$ $(\boldsymbol{x}^\star,\boldsymbol{u}^\star)$ is saddle point of $\mathcal{L}$ $\iff$ $\boldsymbol{x}^\star$ is solution (OPT) and $\boldsymbol{u}^\star$ minimizes $(f+g)^*(-\boldsymbol{u}) + h^*(\boldsymbol{u})$.

Convergence Analysis

Let f, g, h be convex, and f also differentiable with Lipschitz gradient.

We define $s_t \stackrel{\text{def}}{=} \sum_{i=0}^{t-1}\gamma_t$, $\overline{\boldsymbol{x}}_t \stackrel{\text{def}}{=} \left(\sum_{i=0}^{t-1}\gamma_i\boldsymbol{x}_{i+1}\right)/s_t$, $\overline{\boldsymbol{u}}_t \stackrel{\text{def}}{=} \left(\sum_{i=0}^{t-1}\gamma_i\boldsymbol{u}_{i+1}\right)/s_t$.

Theorem (sublinear convergence). For any $(\boldsymbol{x},\boldsymbol{u}) \in \text{dom}\mathcal{L}$:

$$\mathcal{L}(\overline{\boldsymbol{x}}_{t+1},\boldsymbol{u}) - \mathcal{L}(\boldsymbol{x},\overline{\boldsymbol{u}}_{t+1}) \le \frac{\|\boldsymbol{z}_0 - \boldsymbol{x}\|^2 + \gamma_0^2\|\boldsymbol{u}_0 - \boldsymbol{u}\|^2}{2s_t}\,.$$

Corollary (sublinear on objective). If h is β_h-Lipschitz,

$$(f+g+h)(\overline{\boldsymbol{x}}_{t+1}) - (f+g+h)(\boldsymbol{x}^*) \le \frac{\|\boldsymbol{z}_0 - \boldsymbol{x}^*\|^2 + 2\gamma_0^2(\|\boldsymbol{u}_0\|^2 + \beta_h^2)}{2s_t} = \mathcal{O}(1/t)$$

Convergence Analysis

Theorem (linear convergence). If f is μ_f-strongly cvx and h is L_h-smooth,

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \le \left(1 - \min\left\{\tau\frac{\mu_f}{L_f}, \frac{1}{1+\gamma_0 L_h}\right\}\right)^{t+1} C_0\,,$$

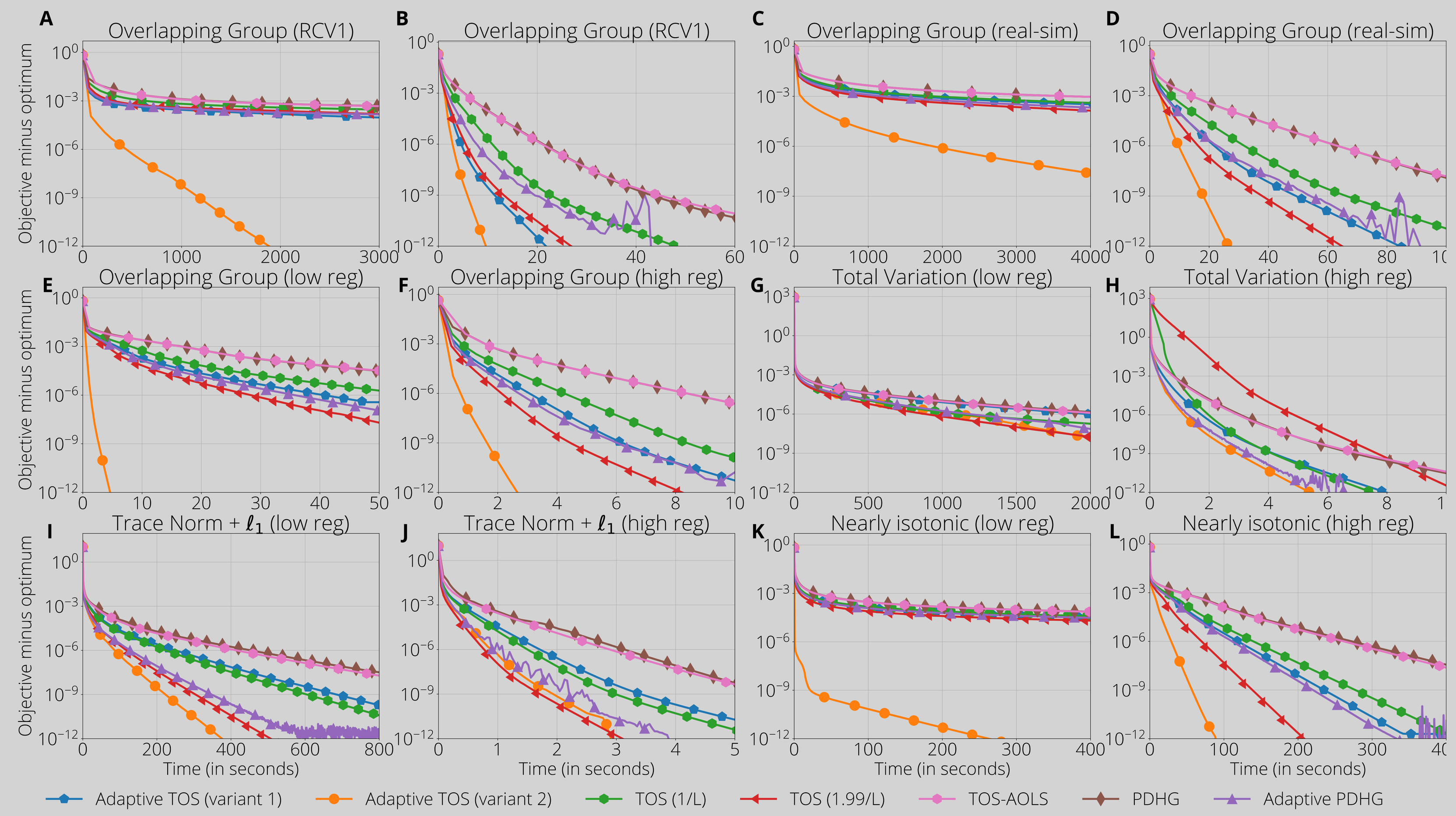
with $C_0 \stackrel{\text{def}}{=} 6\|\boldsymbol{z}_0 - \boldsymbol{x}^*\|^2 + \frac{6}{1-\sigma}\|\gamma_0(\boldsymbol{u}_0 - \boldsymbol{u}^*)\|^2$.

- Rate factor $\min\{\frac{\mu_f}{L_f}, \frac{1}{1+\gamma_0 L_h}\}$ larger than previousl $\frac{\mu_f}{L_f} \times \frac{1}{(1+\gamma_0 L_h)^2}$ (Davis and Yin 2015).

Experiments

Comparison with related splitting methods (Condat 2013, PDHG), (Giselsson et al. 2016, TOS-AOLS), (Malitsky and Pock 2018, Adaptive-PDHG).

Considered problems: logistic regression + nearly-isotonic, overlapping group lasso and least squares + total variation, large and small regularization regime.



Large computational gains with

- non-quadratic los functions
- low regularization regime.

References

Condat, Laurent (2013). "A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms". In: *Journal of Optimization Theory and Applications.*

Davis, Damek and Wotao Yin (2015). "A three-operator splitting scheme and its optimization applications". In: *arXiv preprint arXiv:1504.01032.*

— (2017). "A three-operator splitting scheme and its optimization applications". In: *Set-Valued and Variational Analysis.*

Giselsson, Pontus et al. (2016). "Line search for averaged operator iteration". In: *Conference on Decision and Control (CDC).*

Malitsky, Yura and Thomas Pock (2018). "A first-order primal-dual algorithm with linesearch". In: *SIAM Journal on Optimization.*

http://github.com/openopt/copt