Problem Setting

Goal: solve optimization problems of the form

\[
\min_{x \in \mathbb{R}^d} f(x) + g(x) + h(x),
\]

(OPT)

with access to \(\nabla f\), \(\text{prox}_{g} \) \(\text{arg}\min_{x} g(x) + \frac{1}{2\tau}\|x - z\|^2\), \(\text{prox}_{h}\).

- 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, \(\ell_1\) 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 \(y\):

\[
z = \text{prox}_g(y), \quad x = \text{prox}_g(2y - z - \gamma \nabla f(z)), \quad y = y - z + x
\]

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

\[
\min_x \varphi(x) + \sum_{j=1}^J h_j(x) \equiv \min_x \varphi(x) + \sum_{j=1}^J h_j(x) + \langle X, \cdots, X \rangle.
\]

Convergence has been proven for step-sizes \(\gamma < 2/L\), with \(L\) is Lipschitz constant of \(\nabla f\).

- 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 as 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 \(z\):

\[
Q_f(x, \gamma) \overset{\text{def}}{=} f(z) + \langle \nabla f(z), x - z \rangle + \frac{1}{2\gamma}\|x - z\|^2.
\]

- Choose \(\gamma\) such that \(Q_f(x_{t+1}, \gamma)\) is upper bound of \(f(x_{t+1})\).

Algorithm

1. Start with optimistic step-size \(\gamma_0\) and decrease it until:

\[
f(x_{t+1}) \leq f(z) + \langle \nabla f(z), x_{t+1} - z \rangle + \frac{1}{2\gamma_0}\|x_{t+1} - z\|^2
\]

with \(x_{t+1} = \text{prox}_{\gamma_0} (z - \gamma_0 \nabla f(z) + u_t)\).

2. Run the following updates selected step-size:

\[z_{t+1} = \text{prox}_{\gamma_0} (z_{t+1} + \gamma_0 u_{t+1}) - \gamma_0 u_{t+1} - (x_{t+1} - z_{t+1})\]

Analysis Framework

\[\Delta\] Iterates are not always feasible, objective can be \(\infty\) \[\Delta\]

A better gap function emerges from the saddle-point formulation

\[
\min_{x \in \mathbb{R}^d} f(x) + g(x) + h(x) = \min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} \langle x, y \rangle - f(y)
\]

- \(\mathcal{L}(x, u) < \infty\)
- \(\min_u \mathcal{L}(x, u) = f(x) + g(x) + h(x)\)
- \(\mathcal{L}(x^*, u^*) \leq 0\) for all \((x, u) \iff x^*\), \(u^*\) is saddle point of \(\mathcal{L} \iff x^*\) is solution (OPT) and \(u^*\) minimizes \((f + g)^*(\cdot - u^*) + h^*(u)\).

Convergence Analysis

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

We define \(\alpha_i = \sum_{j=1}^T g_j \) \(\overline{x}_t = \sum_{j=1}^T g_j x_{t+1} / \alpha_t\)

\[\overline{u}_t = \sum_{j=1}^T g_j u_{t+1} / \alpha_t\]

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

\[
\mathcal{L}(x_{t+1}, u_{t+1}) - \mathcal{L}(x, u) \leq \frac{\|x_0 - x\|^2 + \|u_0 - u\|^2}{2\alpha_t}
\]

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

\[
(f + g + h)(x_{t+1}) - (f + g + h)(x^*) \leq \frac{\|x_0 - x^\|^2 + \|u_0 - u^\|^2 + L_h^2}{2\alpha_t} = O(1/t)
\]

Convergence Analysis

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

\[
\|x_{t+1} - x^*\|^2 \leq \left(1 - \frac{\mu}{2L_h + \mu}\right)^t \|x_0 - x^*\|^2
\]

with \(C_0 = \frac{1}{\gamma^2}\|x_0 - x^*\|^2 + \frac{1}{\gamma^2}\|u_0 - u^*\|^2\).

- Rate factor \(\min \left\{ \frac{\gamma^2}{L_h + \gamma} \right\} \) larger than previal \(\gamma^2/(L_h + \gamma)\) (Davis and Yin 2015).

Experiments


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 loss functions
- low regularization regime.

References


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


http://github.com/opento/copt