Stochastic Decoupling Method

Konstantin Mishchenko
Work done together with Peter Richtárik

جامعة الملك عبدالله
للعلوم والتقنية
King Abdullah University of
Science and Technology

Plan

1. Problem structure
2. Examples
3. Proposed method
4. Convergence rates
5. Experiments

Plan

1. Problem structure

Structure

$$\min_x f(x) + \sum_{j=1}^{m} g_j(x)$$

Convex

Structure

$$\min_x f(x) + \sum_{j=1}^{m} g_j(x)$$

Differentiable and smooth

Structure

$$\min_x f(x) + \sum_{j=1}^{m} g_j(x)$$

Proximable

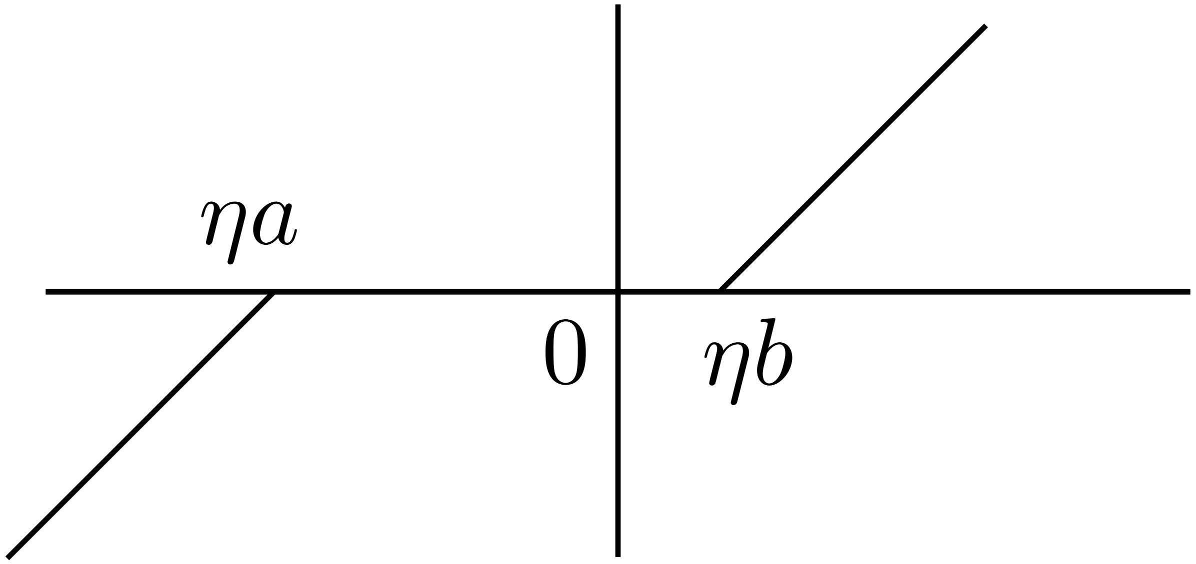
$$\mathrm{prox}_{\eta g}(x) \stackrel{def}{=} \arg\min_u \left\{ g(u) + \frac{1}{2\eta}\|u - x\|^2 \right\}$$

Structure

Proximable

$$g_j = \begin{cases} ax, \ x < 0, \\ bx, \ x \geq 0 \end{cases}$$

$$\mathrm{prox}_{\eta g_j}(x) = \begin{cases} x - \eta a, \ x < \eta a \\ 0, \ \eta a \leq x \leq \eta b \\ x - \eta b, \ \text{otherwise} \end{cases}$$



Plan

2. Examples

Examples

$$\min_x f(x) + \sum_{j=1}^m g_j(x)$$

$$\min_x \frac{1}{2}\|x - x^0\|^2 + \sum_{j=1}^m \chi_{\{z : a_j^\top z = b_j\}}(x)$$

Examples

$$\min_x f(x) + \sum_{j=1}^m g_j(x)$$

$$\min_x \frac{1}{2}\|x - x^0\|^2 + \sum_{j=1}^m \chi_{\{z : a_j^\top z = b_j\}}(x)$$

$$\chi_C(x) = \begin{cases} 0, & x \in C, \\ +\infty, & x \notin C \end{cases}$$

Examples

$$\min_x \left\{ \|x - x^0\| \mid \mathbf{A}x = b \right\}$$

$$j \sim U(1, \ldots, n)$$

$$x^{k+1} = \Pi_j(x^k)$$

Kaczmarz Algorithm, 1937

Examples

$$\min_x \left\{ \|x - x^0\| \mid \mathbf{A}x = b \right\}$$

$$j \sim U(1, \ldots, n)$$

$$x^{k+1} = \Pi_j(x^k)$$

Kaczmarz Algorithm, 1937

Examples

$$\min_x \{\|x - x^0\| \mid \mathbf{A}x = b\}$$

$$j \sim U(1, \ldots, n)$$

$$x^{k+1} = \Pi_j(x^k)$$

Kaczmarz Algorithm, 1937

… revisited in 2009

Examples

$$\min_x f(x) + \sum_{j=1}^{m} g_j(x)$$

$$\min_x \{f(x) \mid \mathbf{C}x = d\}$$

$$\min_x \left\{ \frac{1}{n} \sum_{i=1}^{n} f_i(x_i) \mid x_1 = x_2 = \cdots = x_n \right\}$$

Examples

$$\min_x f(x) + \sum_{j=1}^{m} g_j(x)$$

$$\min_x f(x) + \|\mathbf{B}x\|_1$$

Examples

$$\min_x f(x) + \sum_{j=1}^{m} g_j(x)$$

$$\min_x f(x) + \|\mathbf{B}x\|_1$$

For instance, Fused LASSO (Tibshirani et al., 2005)

$$b_j = (0, 0, \ldots, \underbrace{1, -1}_{j,j+1}, 0, \ldots, 0)$$

Examples

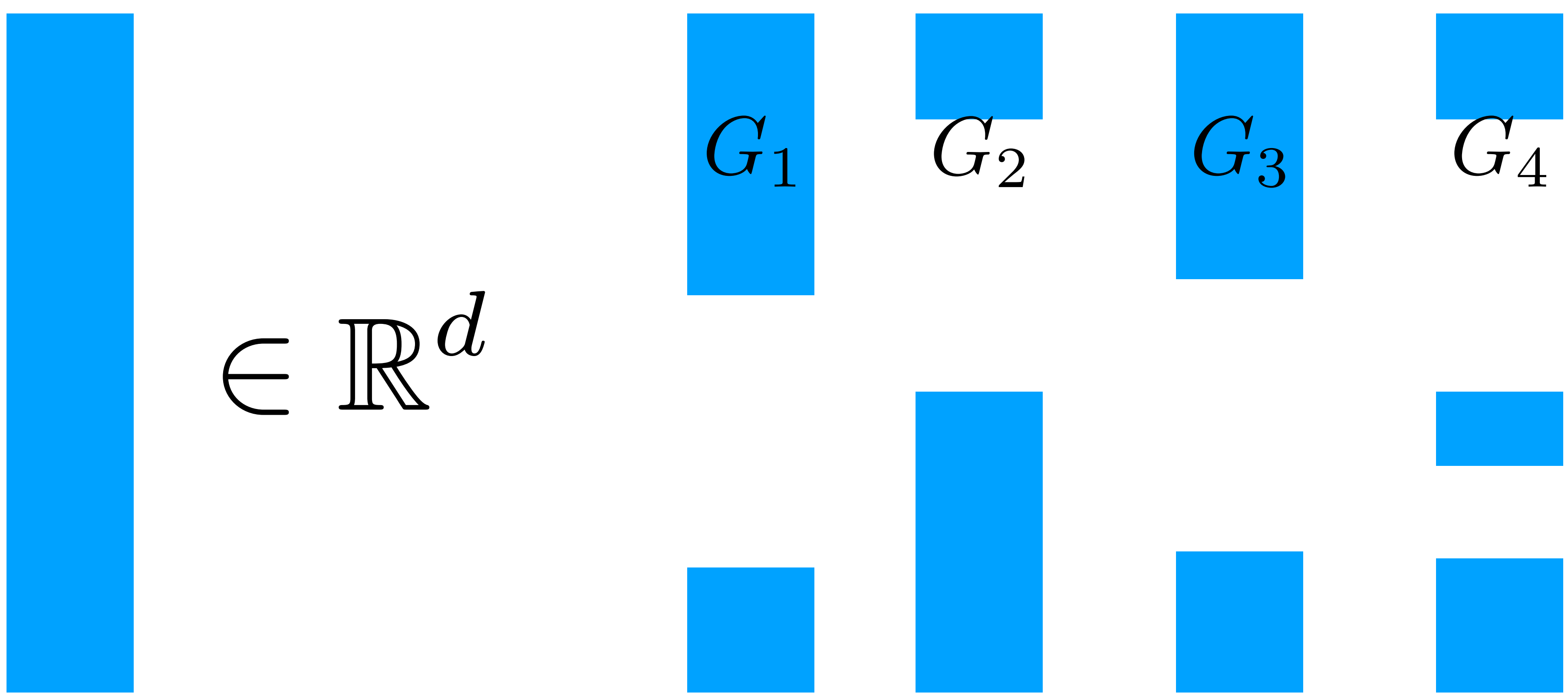
$$\min_x f(x) + \sum_{j=1}^{m} g_j(x)$$

$$\min_x f(x) \ \text{s.t.} \ x \in \bigcap_{j=1}^{m} C_j$$

Examples

$$\min_x f(x) + \sum_{j=1}^{m} g_j(x)$$

$$\min_x \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \boxed{\sum_{j=1}^{m} \|x\|_{G_j}}$$



Examples

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

$$\min_x \frac{1}{2}x^\top \mathbf{A}x + b^\top x$$

$$\min_x \frac{1}{n}\sum_{i=1}^{n} \log\left(1 + \exp(-b_i a_i^\top x)\right)$$

$$\min_x \frac{1}{n}\sum_{i=1}^{n} l\left(b_i, \Phi(x, a_i)\right) \qquad a_i \in \mathbb{R}^{d_1}, b_i \in \mathbb{R}$$

Examples

$$\min_x f(x) + \sum_{j=1}^{m} g_j(x)$$

$$\frac{1}{2}\|y - \mathbf{A}x\|_2^2 + \lambda\|x\|_1 + \lambda_1 \sum_{j=1}^{m} \|\mathbf{R}_j x\|_2$$

“We have not experimented with this yet, as the computation seems challenging due to the presence of ℓ_2 norms.” (Tay, Friedman, Tibshirani, PCA-Lasso 2018)

Plan

3. Proposed method

Gradient descent

$$x^{t+1} = x^t - \eta \nabla f(x^t), \quad t = 0, 1 \ldots, T$$

Proximal gradient descent

$$x^{t+1} = \mathrm{prox}_{\eta g}(x^t - \eta \nabla f(x^t))$$

Proximal gradient descent

$$x^{t+1} = \boxed{\operatorname{prox}_{\eta g}}(x^t - \eta \nabla f(x^t))$$

$$\operatorname{prox}_{\eta g}(x) \stackrel{def}{=} \arg\min_u \left\{ g(u) + \frac{1}{2\eta} \|u - x\|^2 \right\}$$

Gradient descent

$$x^{t+1} = \mathrm{prox}_{\eta g}(x^t - \eta \nabla f(x^t))$$

$$f(x^t) - \min_x f(x) = \boxed{\mathcal{O}\left(\left(1 - \frac{\mu}{L}\right)^t\right)}$$

Linear rate

Stochastic decoupling

$$y^t = \frac{1}{m}\sum_{j=1}^{m} y_j^t$$

$$z^t = x^t - \boxed{\eta \nabla f(x^t)} - \eta y^t$$

Stochastic decoupling

$$y^t = \frac{1}{m}\sum_{j=1}^{m} y_j^t$$

$$z^t = x^t - \eta \nabla f(x^t) \boxed{- \eta y^t}$$

$$y_j^t \approx \partial g_j(x^t), \quad y^t \approx \partial g(x^t)$$

$$\boxed{z^t \approx \mathrm{prox}_{\eta g}(x^t - \eta \nabla f(x^t))}$$

Stochastic decoupling

$$y^t = \frac{1}{m}\sum_{j=1}^{m} y_j^t$$

$$z^t = x^t - \eta\nabla f(x^t) - \eta y^t$$

$$y_j^t \approx \partial g_j(x^t), \quad \boxed{y^t \approx \partial g(x^t)}$$

$$z^t \approx \mathrm{prox}_{\eta g}(x^t - \eta\nabla f(x^t))$$

$$x^{t+1} = \mathrm{prox}_{\eta g_j}(z^t + \eta y_j^t)$$

$$y_j^{t+1} = y_j^t + \frac{1}{\eta}(z^t - x^{t+1}) \boxed{\in \partial g_j(x^{t+1})}$$

Plan

4. Convergence rates

Convergence

$\mathcal{O}(1/\varepsilon)$ **Convex**

Convergence

$\mathcal{O}(1/\varepsilon)$ Convex

$\mathcal{O}(1/\sqrt{\varepsilon})$ f is μ-strongly convex

Convergence

$\mathcal{O}(1/\varepsilon)$ Convex

$\mathcal{O}(1/\sqrt{\varepsilon})$ f is μ-strongly convex

$\mathcal{O}(\log \frac{1}{\varepsilon})$ $g_j(x) = \phi_j(a_j^\top x)$

$\mathbf{A}^\top \mathbf{A} \succ 0$

Convergence

$\mathcal{O}(1/\varepsilon)$ Convex

$\mathcal{O}(1/\sqrt{\varepsilon})$ f is μ-strongly convex

$\mathcal{O}(\log \frac{1}{\varepsilon})$ $g_j(x) = \phi_j(a_j^\top x)$

$\mathbf{A}^\top \mathbf{A} \succ 0$

Was only possible for $f = \frac{1}{2}\|x - x^0\|^2$

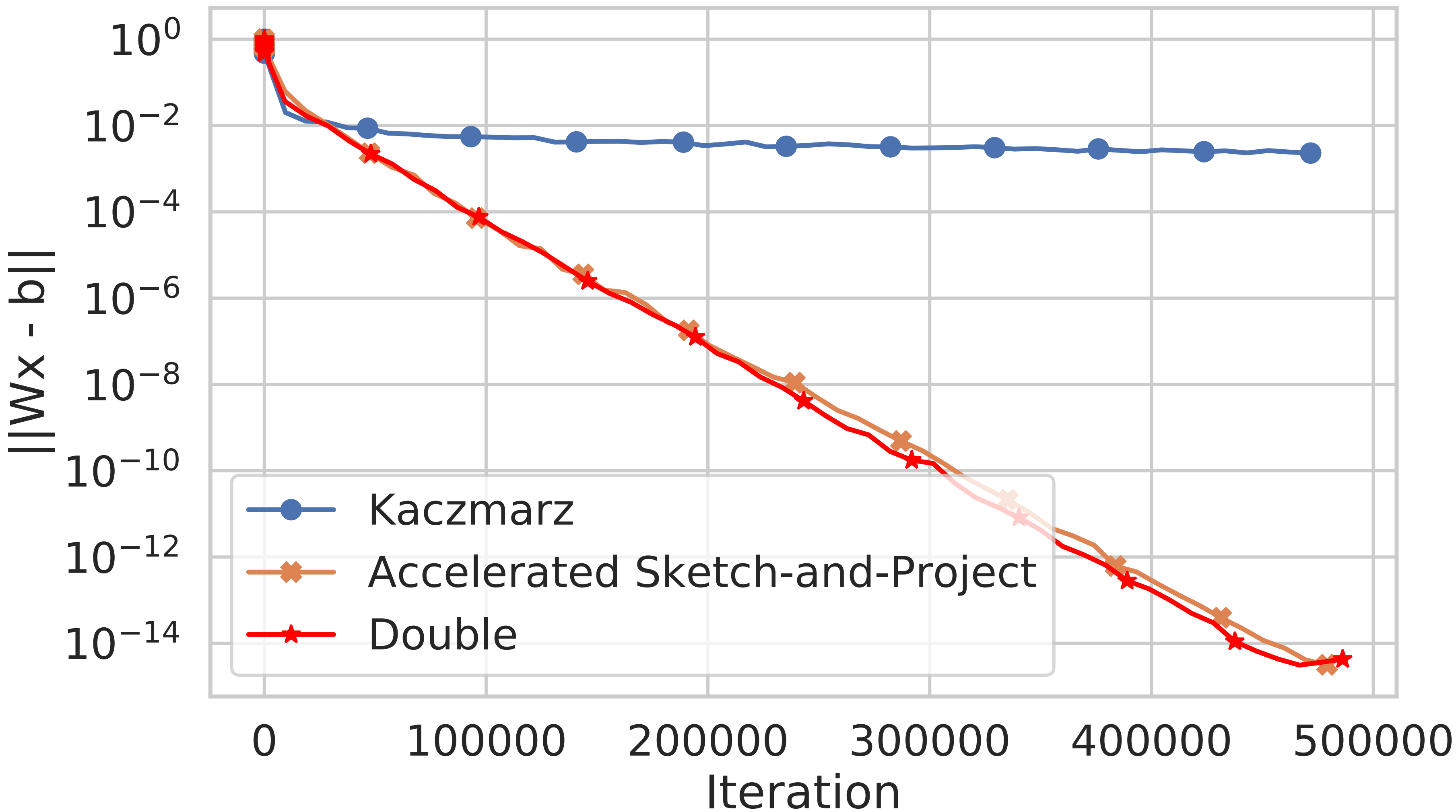
before our work

Plan

5. Experiments

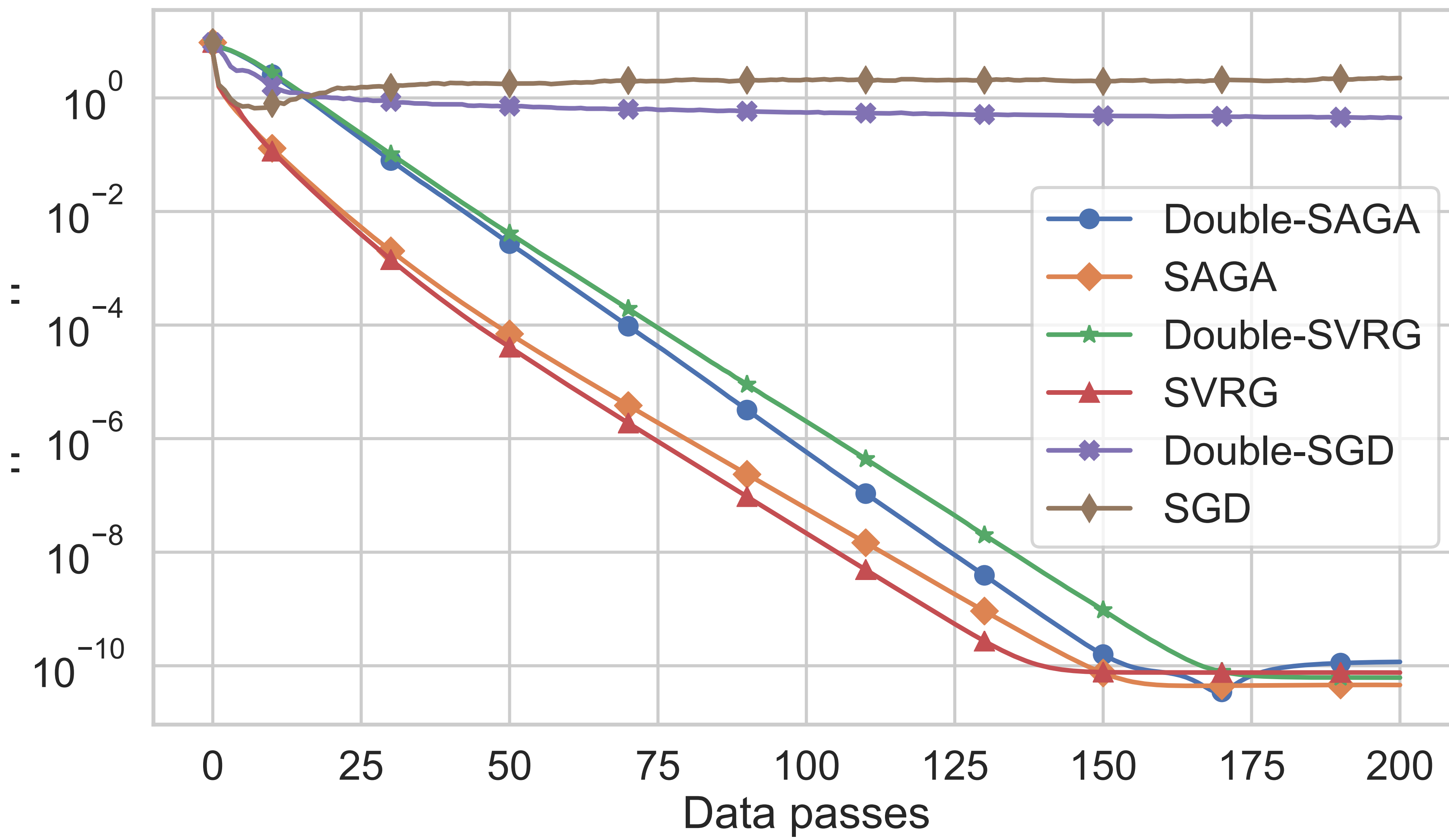
Experiments

$$\min_x \{\|x - x^0\| \mid \mathbf{W}x = b\}$$



Experiments

$$\min_x \left\{ \frac{1}{2} x^\top \mathbf{A} x + b^\top x \mid \mathbf{C}x = d \right\}$$



Reference

A Stochastic Decoupling Method for Minimizing the Sum of Smooth and Non-Smooth Function

arXiv:1905.11535