Universal first-order methods for convex and strongly convex hybrid composite optimization

Jiaming Liang

Department of Computer Science & Goergen Institute for Data Science
University of Rochester

July 23, 2024

Joint work with Vincent Guigues (FGV) and Renato Monteiro (Georgia Tech)

International Symposium on Mathematical Programming, Montréal, Canada

Introduction

Consider convex hybrid composite optimization (HCO) problem

$$\phi_* := \min \{\phi(x) := f(x) + h(x) : x \in \mathbb{R}^n\},$$

where $f, h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are proper lower semi-continuous convex functions and h has a simple proximal mapping.

Complexities of first-order methods for obtaining an ε-solution to HCO:

- f is L-smooth, i.e., $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$, for every $x, y \in \mathrm{dom}\, h$, by the Nesterov's accelerated gradient method, $\mathcal{O}(\sqrt{L} d_0/\sqrt{\varepsilon})$
- f is M-Lipschitz continuous, i.e., $\|f'(x)\| \le M$, for every $x \in \mathrm{dom}\, h$, by the subgradient method, $\mathcal{O}(M^2 d_0^2/\varepsilon^2)$

What if for some $\alpha \in (0, 1)$, f has α-Hölder continuous gradient, namely, for every $x, y \in \mathrm{dom}\, h$, $\|\nabla f(x) - \nabla f(y)\| \le L_\alpha \|x - y\|^\alpha$?

Universal methods

- Universal fast gradient in (Nesterov, 2015)
- Accelerated bundle-level and accelerated prox-level in (Lan, 2015)
- Without knowing/using any of the parameters α and L_α, the above methods have complexity

$$\tilde{\mathcal{O}}\left(\left(\frac{d_0^{1+\alpha}L_\alpha}{\bar{\varepsilon}}\right)^{\frac{2}{1+3\alpha}}\right)$$

- Universal primal gradient (UPG) in (Nesterov, 2015), which is an adaptive subgradient method, has complexity

$$\tilde{\mathcal{O}}\left(\left(\frac{d_0^{1+\alpha}L_\alpha}{\bar{\varepsilon}}\right)^{\frac{2}{1+\alpha}}\right)$$

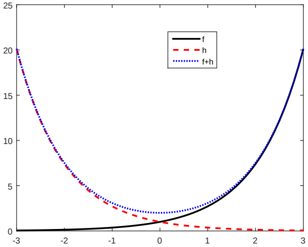
What if $\phi = f + h$ is μ-strongly convex? What is the complexity of UPG?

μ-universal methods

We are interested in parameter-free methods whose complexities for solving HCO are expressed in terms of μ and term them as μ-universal method.

(Liang and Monteiro, 2024) shows that UPG is μ_h-universal, but it is not known if it is μ-universal.

μ can be substantially larger than $\mu_f + \mu_h$ (e.g., for $\alpha \gg 0$, $f(x) = \alpha \exp(x)$, and $h(x) = \alpha \exp(-x)$, we have $\mu = 2\alpha \gg 0 = \mu_f + \mu_h$)



By a slight generalization of UPG, we can show it is μ-universal.

Comparison with μ-universal methods

Smooth

- Papers concerned with function values
  - restart based on estimate of μ: (Nesterov, 2013) and (Fercoq and Qu, 2019)
  - assuming ϕ_* is known: (Renegar and Grimmer, 2022)
- Papers concerned with stationary points
  - restart based on estimate of μ: (Alamo, Krupa, and Limon, 2019), (Aujol, Dossal, Labarrière, and Rondepierre, 2022), and (Lan, Ouyang, and Zhang, 2023)
  - assuming ϕ_* is known: (Aujol, Dossal, and Rondepierre, 2023)

Non-smooth

Assuming ϕ_* is known: (Renegar and Grimmer, 2022) and (Grimmer, 2023)

In contrast, our generalization of UPG is μ-universal. Both functional and stationary complexity bounds. No restart scheme, no prior knowledge of ϕ_*.

Overview

- We present two μ-universal methods: universal composite subgradient (U-CS) and univesal proximal bundle (U-PB)

- We establish both functional and stationary complexity bounds for U-CS and U-PB

- Both methods are analyzed in a unified manner using a general framework for strongly convex optimization, denoted by FSCO

- No line-search/restart based on estimate of μ

Outline

1. Framework for strongly convex optimization
2. Universal composite subgradient
3. Universal proximal bundle

Outline

1. Framework for strongly convex optimization
2. Universal composite subgradient
3. Universal proximal bundle

FSCO

FSCO is presented in the context of strongly convex optimization problem

$$\phi_* := \min\{\phi(x) : x \in \mathbb{R}^n\}$$

where $\phi \in \overline{\text{Conv}}_\mu(\mathbb{R}^n)$, i.e., ϕ is μ-strongly convex.

Algorithm FSCO

---

1. Let $\chi \in [0,1)$, $\varepsilon > 0$, and $\hat{x}_0 \in \text{dom}\,\phi$ be given, and set $k = 1$;
2. Compute $\lambda_k > 0$, $\hat{\Gamma}_k \in \overline{\text{Conv}}(\mathbb{R}^n)$, $\hat{\Gamma}_k \le \phi$, and $\hat{y}_k \in \text{dom}\,\phi$ satisfying

$$\phi(\hat{y}_k) + \frac{\chi}{2\lambda_k}\|\hat{y}_k - \hat{x}_{k-1}\|^2 - \min_{u\in\mathbb{R}^n}\left\{\hat{\Gamma}_k(u) + \frac{1}{2\lambda_k}\|u - \hat{x}_{k-1}\|^2\right\} \le \varepsilon,$$

   and set
$$\hat{x}_k := \underset{u\in\mathbb{R}^n}{\text{argmin}}\left\{\hat{\Gamma}_k(u) + \frac{1}{2\lambda_k}\|u - \hat{x}_{k-1}\|^2\right\};$$

3. Check whether a termination criterion holds and if so **stop**; else go to step 4;
4. Set $k \leftarrow k+1$ and go to step 2.

Complexity analysis – function value

Assumptions

(F1) there exists $\nu \in [0, \mu]$ such that $\hat{\Gamma}_k \in \overline{\text{Conv}}_\nu(\mathbb{R}^n)$;

(F2) there exists $\underline{\lambda} > 0$ such that $\lambda_k \geq \underline{\lambda}$ for every iteration k of the FSCO.

Theorem

For a given tolerance $\bar{\varepsilon} > 0$, consider FSCO with $\varepsilon = (1-\chi)\bar{\varepsilon}/2$, where $\chi \in [0, 1)$. Then, the number of iterations of FSCO to generate a $\bar{\varepsilon}$-solution is at most

$$\mathcal{C}_{func}(\bar{\varepsilon}) := \min\left\{\min\left[\frac{1}{\chi}\left(1+\frac{1}{\underline{\lambda}\mu}\right), 1+\frac{1}{\underline{\lambda}\nu}\right]\log\left(1+\frac{\lambda_0\mu d_0^2}{\underline{\lambda}\bar{\varepsilon}}\right), \frac{d_0^2}{\underline{\lambda}\bar{\varepsilon}}\right\}.$$

Complexity analysis – stationary point

Theorem

For a given tolerance pair $(\hat{\varepsilon}, \hat{\rho}) \in \mathbb{R}^2_{++}$, *FSCO with*

$$\chi \in (0,1), \quad \varepsilon = \frac{\chi(1-\chi)\hat{\varepsilon}}{10},$$

generates a triple $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ *satisfying*

$$\bar{s}_k \in \partial \phi_{\bar{\varepsilon}_k}(\bar{y}_k), \quad \|\bar{s}_k\| \le \hat{\rho}, \quad \bar{\varepsilon}_k \le \hat{\varepsilon}$$

in at most

$$\min\left\{ \min\left[\frac{1}{\chi}\left(1 + \frac{1}{\underline{\lambda}\mu}\right), 1 + \frac{1}{\underline{\lambda}\nu} \right] \log\left[1 + \lambda_0 \mu \beta(\hat{\varepsilon}, \hat{\rho})\right], \beta(\hat{\varepsilon}, \hat{\rho}) \right\}$$

iterations where

$$\beta(\hat{\varepsilon}, \hat{\rho}) = \frac{1}{\underline{\lambda}}\left(\frac{4\chi\hat{\varepsilon}}{5\hat{\rho}^2} + \frac{5 d_0^2}{\chi\hat{\varepsilon}} \right).$$

Outline

U-CS

Assumptions

(A1) $h \in \overline{\text{Conv}}_\nu(\mathbb{R}^n)$ for some $0 \le \nu \le \mu$;

(A2) $f \in \overline{\text{Conv}}(\mathbb{R}^n)$ is such that $\text{dom}\, h \subset \text{dom}\, f$, and has a subgradient oracle;

(A3) there exists $(M_f, L_f) \in \mathbb{R}^2_+$ such that for every $x, y \in \text{dom}\, h$,

$$\|f'(x) - f'(y)\| \le 2M_f + L_f\|x - y\|.$$

α-Hölder continuous gradient implies (A3) (Liang and Monteiro, 2024)

Algorithm U-CS

1. Let $\hat{x}_0 \in \text{dom}\, h$, $\chi \in [0, 1)$, $\lambda_0 > 0$, and $\varepsilon > 0$ be given, and set $\lambda = \lambda_0$ and $j = 1$;
2. Compute
$$x = \underset{u \in \mathbb{R}^n}{\text{argmin}} \left\{ \ell_f(u; \hat{x}_{j-1}) + h(u) + \frac{1}{2\lambda}\|u - \hat{x}_{j-1}\|^2 \right\};$$
3. **If** $f(x) - \ell_f(x; \hat{x}_{j-1}) - (1 - \chi)\|x - \hat{x}_{j-1}\|^2/(2\lambda) \le \varepsilon$ does not hold, **then** set $\lambda = \lambda/2$ and go to step 2; **else**, go to step 4;
4. Set $\lambda_j = \lambda$, $\hat{x}_j = x$, $j \leftarrow j + 1$, and go to step 2.

U-CS as an instance of FSCO

Proposition

The following statements hold for U-CS:

a) $\{\lambda_k\}$ *is a non-increasing sequence;*

b) *for every* $k \geq 1$, *we have*

$$\hat{x}_k = \underset{u \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \ell_f(u; \hat{x}_{k-1}) + h(u) + \frac{1}{2\lambda_k} \|u - \hat{x}_{k-1}\|^2 \right\},$$

$$f(\hat{x}_k) - \ell_f(\hat{x}_k; \hat{x}_{k-1}) + \frac{\chi - 1}{2\lambda_k} \|\hat{x}_k - \hat{x}_{k-1}\|^2 \leq \varepsilon,$$

$$\lambda_k \geq \underline{\lambda}(\varepsilon) := \min \left\{ \frac{(1-\chi)\varepsilon}{4\left(\overline{M}_f^2 + \varepsilon \overline{L}_f\right)}, \lambda_0 \right\}.$$

c) *U-CS is a special case of FSCO where:*

   i) $\hat{y}_k = \hat{x}_k$ *and* $\hat{\Gamma}_k(\cdot) = \ell_f(\cdot; \hat{x}_{k-1}) + h(\cdot)$ *for every* $k \geq 1$;

   ii) *assumptions (F1) and (F2) are satisfied with* $\underline{\lambda} = \underline{\lambda}(\varepsilon)$ *and* ν *from assumption (A3).*

Complexity analysis – function value

Theorem

Let $\bar{\varepsilon} > 0$ be given and consider U-CS with $\varepsilon = (1-\chi)\bar{\varepsilon}/2$, where $\chi \in [0,1)$ is as in step 0 of U-CS. Then, the number of iterations of U-CS to generate an iterate $\hat{x}_k$ satisfying $\phi(\hat{x}_k) - \phi_* \le \bar{\varepsilon}$ is at most

$$\min\left\{\min\left[\frac{1}{\chi}\left(1+\frac{Q_f(\bar{\varepsilon})}{\mu\bar{\varepsilon}}\right), 1+\frac{Q_f(\bar{\varepsilon})}{\nu\bar{\varepsilon}}\right]\log\left(1+\frac{\lambda_0\mu Q_f(\bar{\varepsilon})d_0^2}{\bar{\varepsilon}^2}\right), \frac{d_0^2 Q_f(\bar{\varepsilon})}{\bar{\varepsilon}^2}\right\} + \left\lceil 2\log\frac{\lambda_0 Q_f(\bar{\varepsilon})}{\bar{\varepsilon}}\right\rceil$$

where

$$Q_f(\bar{\varepsilon}) = \frac{8\overline{M}_f^2}{(1-\chi)^2} + \bar{\varepsilon}\left(\lambda_0^{-1} + \frac{8\overline{L}_f}{(1-\chi)^2}\right)$$

U-CS: $\tilde{\mathcal{O}}(d_0^2(\overline{M}_f^2 + \bar{\varepsilon}\overline{L}_f)/\bar{\varepsilon}^2)$ UPG: $\tilde{\mathcal{O}}\left(d_0^2(L_\alpha/\bar{\varepsilon})^{\frac{2}{\alpha+1}}\right)$

By (Liang and Monteiro, 2024), we know $\overline{M}_f^2 + \bar{\varepsilon}\overline{L}_f \le 2\bar{\varepsilon}^{\frac{2\alpha}{\alpha+1}} L_\alpha^{\frac{2}{\alpha+1}}$, so U-CS has a better complexity.

Complexity analysis – stationary point

Recall a $(\hat{\rho}, \hat{\varepsilon})$-stationary solution is a triple $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ such that

$$\bar{s}_k \in \partial \phi_{\bar{\varepsilon}_k}(\bar{y}_k), \quad \|\bar{s}_k\| \le \hat{\rho}, \quad \bar{\varepsilon}_k \le \hat{\varepsilon}.$$

For U-CS, we define

$$\bar{y}_k = \operatorname{argmin}\{\phi(y) : y \in \{\hat{x}_1, \ldots, \hat{x}_k\}\},$$
$$\bar{s}_k = \frac{\hat{x}_0 - \hat{x}_k}{S_k}, \quad \bar{\varepsilon}_k = \frac{\|\hat{x}_0 - \bar{y}_k\|^2 - \|\hat{x}_k - \bar{y}_k\|^2}{2S_k} + \frac{\varepsilon}{1-\chi}.$$

Complexity analysis – stationary point

Theorem

For a given tolerance pair $(\hat{\rho}, \hat{\varepsilon}) \in \mathbb{R}^2_{++}$*, consider U-CS with* $\chi \in [0,1)$ *and* $\varepsilon = (1-\chi)\hat{\varepsilon}/6$*. Then for every* $k \geq 1$*, U-CS generates a* $(\hat{\rho}, \hat{\varepsilon})$*-stationary solution* $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ *within a number of iterations bounded by*

$$\min\left\{\min\left[\frac{1}{\chi}\left(1+\frac{Q_s(\hat{\varepsilon})}{\mu\hat{\varepsilon}}\right), 1+\frac{Q_s(\hat{\varepsilon})}{\nu\hat{\varepsilon}}\right]\log C(\hat{\varepsilon},\hat{\rho}),\ \frac{4Q_s(\hat{\varepsilon})}{\hat{\varepsilon}}\left(\frac{\hat{\varepsilon}}{3\hat{\rho}^2}+\frac{d_0^2}{\hat{\varepsilon}}\right)\right\} + \left\lceil 2\log\frac{\lambda_0 Q_s(\hat{\varepsilon})}{\hat{\varepsilon}}\right\rceil$$

where

$$C(\hat{\varepsilon},\hat{\rho}) = 1+\frac{4\lambda_0\mu Q_s(\hat{\varepsilon})}{\hat{\varepsilon}}\left(\frac{\hat{\varepsilon}}{3\hat{\rho}^2}+\frac{d_0^2}{\hat{\varepsilon}}\right), \quad Q_s(\hat{\varepsilon}) = \frac{24\overline{M}_f^2}{(1-\chi)^2}+\hat{\varepsilon}\left(\frac{1}{\lambda_0}+\frac{24\overline{L}_f}{(1-\chi)^2}\right).$$

Outline

Review of the proximal bundle method

Approximately solve the proximal problem

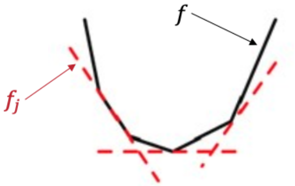
$$\hat{x} := \operatorname{argmin}\left\{f(x) + h(x) + \frac{1}{2\lambda}\|x - x^c\|^2\right\}$$

by an iterative process

$$x_j \leftarrow \min\left\{f_j(x) + h(x) + \frac{1}{2\lambda}\|x - x^c\|^2\right\}.$$

Recursively build up a cutting-plane model

$$f_j(x) = \max_{0\le i\le j-1}\{\ell_f(x;x_i) := f(x_i) + \langle f'(x_i), x - x_i\rangle\}$$



Algorithm U-PB

1. Let $\hat{x}_0 \in \operatorname{dom} h$, $\lambda_1 = \lambda > 0$, $\chi \in [0,1)$, $\varepsilon > 0$, and integer $\overline{N} \geq 1$ be given, and set $y_0 = \hat{x}_0$, $N = 0$, $j = 1$, and $k = 1$. Find $f_1 \in \overline{\text{Conv}}(\mathbb{R}^n)$ such that $\ell_f(\cdot;\hat{x}_0) \leq f_1 \leq f$;
2. Compute
$$x_j = \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ (f_j + h)(u) + \frac{1}{2\lambda}\|u - \hat{x}_{k-1}\|^2 \right\};$$
3. Choose $y_j \in \{x_j, y_{j-1}\}$ such that
$$y_j = \operatorname{argmin} \left\{ \phi(x) + \frac{\chi}{2\lambda}\|x - \hat{x}_{k-1}\|^2 : x \in \{x_j, y_{j-1}\} \right\},$$
and set $N = N + 1$ and
$$t_j = \phi(y_j) + \frac{\chi}{2\lambda}\|y_j - \hat{x}_{k-1}\|^2 - \left((f_j + h)(x_j) + \frac{1}{2\lambda}\|x_j - \hat{x}_{k-1}\|^2 \right);$$

U-PB (continued)

Algorithm U-PB (continued)

4. **If** $t_j > \varepsilon$ and $N < \overline{N}$ **then**
   perform a **null update**, i.e.: set $f_{j+1} = \mathsf{BU}(\hat{x}_{k-1}, x_j, f_j, \lambda)$;
   else
   if $t_j > \varepsilon$ and $N = \overline{N}$
   perform a **reset update**, i.e., set $\lambda \leftarrow \lambda/2$;
   else (i.e., $t_j \leq \varepsilon$ and $N \leq \overline{N}$)
   perform a **serious update**, i.e., set $\hat{x}_k = x_j$, $\hat{\Gamma}_k = f_j + h$, $\hat{y}_k = y_j$, $\lambda_k = \lambda$, and $k \leftarrow k+1$;
   end if
   set $N = 0$ and find $f_{j+1} \in \overline{\mathsf{Conv}}(\mathbb{R}^n)$ such that $\ell_f(\cdot; \hat{x}_{k-1}) \leq f_{j+1} \leq f$;
   end if
5. Set $j \leftarrow j+1$ and go to step 2.

Note: iteration limit $\overline{N}$ and adaptive scheme in λ.

Complexity analysis – functional value

Similarly to U-CS, U-PB can be shown as another instance of FSCO.

Theorem

Given tolerance $\bar{\varepsilon} > 0$*, consider U-PB and* $\varepsilon = (1-\chi)\bar{\varepsilon}/2$*, where* $\chi \in [0,1)$ *is as in step 0 of U-PB. Let* $\{\hat{x}_k\}$ *and* $\{\hat{y}_k\}$ *be the sequences generated by U-PB. Then, the number of iterations of U-PB to generate an iterate* $\hat{y}_k$ *satisfying* $\phi(\hat{y}_k) - \phi_* \le \bar{\varepsilon}$ *is at most*

$$\min\left\{\min\left[\frac{1}{\chi}\left(\overline{N} + \frac{R_f(\bar{\varepsilon})}{\mu\bar{\varepsilon}}\right), \overline{N} + \frac{R_f(\bar{\varepsilon})}{\nu\bar{\varepsilon}}\right]\log\left(1 + \frac{\lambda_0\mu R_f(\bar{\varepsilon})d_0^2}{\bar{\varepsilon}^2\overline{N}}\right), \frac{d_0^2 R_f(\bar{\varepsilon})}{\bar{\varepsilon}^2\overline{N}}\right\}$$
$$+ \overline{N}\left[2\log\frac{\lambda_0 R_f(\bar{\varepsilon})}{\bar{\varepsilon}\overline{N}}\right]$$

where

$$R_f(\bar{\varepsilon}) = \bar{\varepsilon}\overline{N}\left[\frac{1}{\lambda_0} + \frac{40\overline{L}_f}{1-\chi}\right] + \frac{64\overline{M}_f^2}{(1-\chi)^2}\left(1 + \log(\overline{N})\right).$$

Complexity analysis – stationary point

Theorem

For a given tolerance pair $(\hat{\rho}, \hat{\varepsilon}) \in \mathbb{R}^2_{++}$*, U-PB with*

$$\chi \in (0,1), \quad \varepsilon = \frac{\chi(1-\chi)\hat{\varepsilon}}{10},$$

generates a $(\hat{\rho}, \hat{\varepsilon})$*-stationary solution* $(\bar{y}_k, \bar{s}_k, \bar{\varepsilon}_k)$ *in at most*

$$\min\left\{\min\left[\frac{1}{\chi}\left(\overline{N} + \frac{R_s(\hat{\varepsilon})}{\hat{\varepsilon}\mu}\right), \overline{N} + \frac{R_s(\hat{\varepsilon})}{\hat{\varepsilon}\nu}\right]\log C(\hat{\varepsilon}, \hat{\rho}),\ \frac{R_s(\hat{\varepsilon})}{\hat{\varepsilon}\overline{N}}\left(\frac{4\chi\hat{\varepsilon}}{5\hat{\rho}^2} + \frac{5d_0^2}{\chi\hat{\varepsilon}}\right)\right\} + \overline{N}\left[2\log\frac{\lambda_0 R_s(\hat{\varepsilon})}{\hat{\varepsilon}\overline{N}}\right]$$

iterations where

$$C(\hat{\varepsilon}, \hat{\rho}) = 1 + \frac{\lambda_0 \mu R_s(\hat{\varepsilon})}{\hat{\varepsilon}\overline{N}}\left(\frac{4\chi\hat{\varepsilon}}{5\hat{\rho}^2} + \frac{5d_0^2}{\chi\hat{\varepsilon}}\right)$$

$$R_s(\hat{\varepsilon}) = \hat{\varepsilon}\overline{N}\left[\frac{1}{\lambda_0} + \frac{40\overline{L}_f}{1-\chi}\right] + \frac{320\overline{M}_f^2}{\chi(1-\chi)^2}\left(1 + \log(\overline{N})\right).$$

Conclusion

- Two μ-universal methods: U-CS and U-PB
- Both functional and stationary complexities
- Unified analysis through FSCO
- No restart scheme based on estimates of μ

References

- Alamo, Limon, and Krupa. Restart FISTA with global linear convergence. ECC, 2019.
- Aujol, Dossal, and Rondepierre. FISTA is an automatic geometrically optimized algorithm for strongly convex functions. Mathematical Programming, 2023.
- Aujol, Dossal, Labarrière, and Rondepierre. FISTA restart using an automatic estimation of the growth parameter. Preprint, 2022.
- Fercoq and Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. IMA Journal of Numerical Analysis, 2019.
- Grimmer. On optimal universal first-order methods for minimizing heterogeneous sums. Optimization Letters, 2023.
- Lan. Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. Mathematical Programming, 2015.

References

- Lan, Ouyang, and Zhang. Optimal and parameter-free gradient minimization methods for convex and nonconvex optimization. Arxiv preprint 2310.12139, 2023.
- Liang and Monteiro. A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems. Mathematics of Operations Research, 2024.
- **Guigues, Liang, and Monteiro. Universal subgradient and proximal bundle methods for convex and strongly convex hybrid composite optimization. Arxiv preprint 2407.10073, 2024.**
- Nesterov. Gradient methods for minimizing composite functions. Mathematical Programming, 2013.
- Nesterov. Universal gradient methods for convex optimization problems. Mathematical Programming, 2015.
- Renegar and Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. Foundations of Computational Mathematics, 2022.

Thank you!