A Universally Optimal Method for Minimizing Heterogeneously Smooth and Convex Compositions

Aaron Zoll

Department of Applied Math and Statistics
Johns Hopkins University

March 10th 2025

Where do we start?

Compositions?

Heterogeneity?

Smoothness...
Convexity...
Generalized?

Universally
Optimal
Guarantees?

Where do we start?

Compositions?

Heterogeneity?

Smoothness...
Convexity...
Generalized?

Universally
Optimal
Guarantees?

Where do we start?

Compositions?

Heterogeneity?

Smoothness...
Convexity...
Generalized?

Universally
Optimal
Guarantees?

Where do we start?

Compositions?

Heterogeneity?

Smoothness...
Convexity...
Generalized?

Universally
Optimal
Guarantees?

Where do we start?

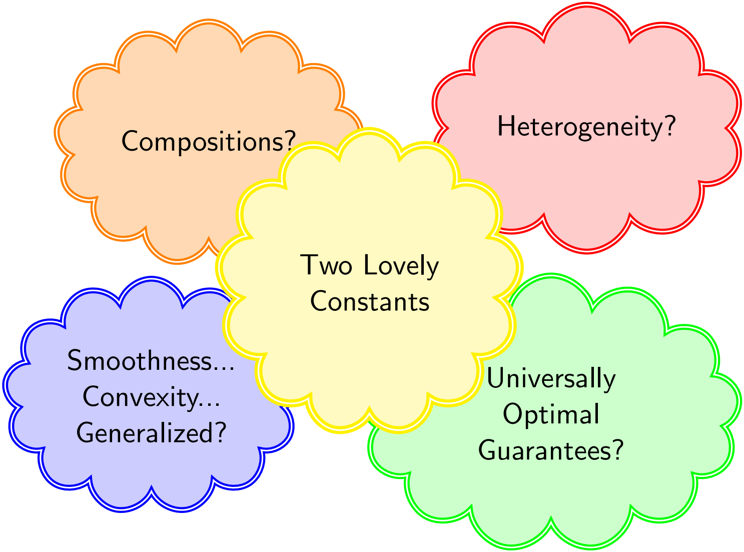
Compositions?

Heterogeneity?

Smoothness...
Convexity...
Generalized?

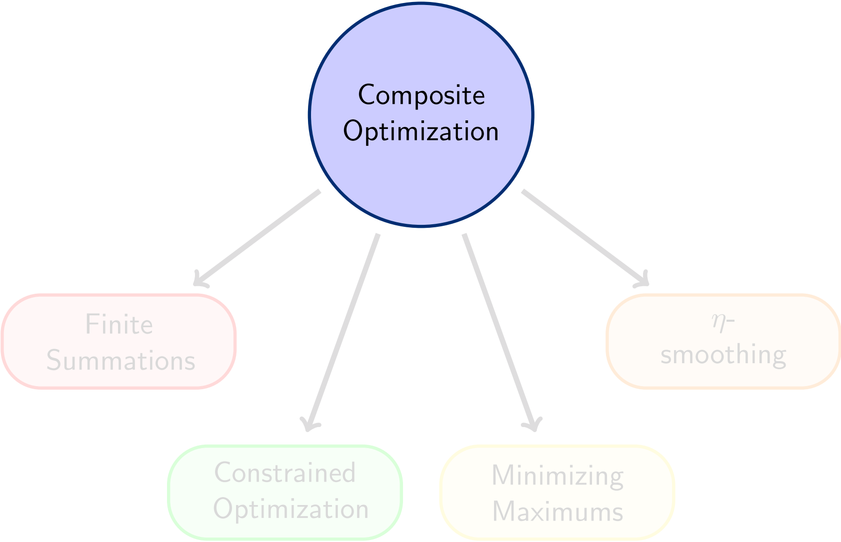
Universally
Optimal
Guarantees?

Where do we start?

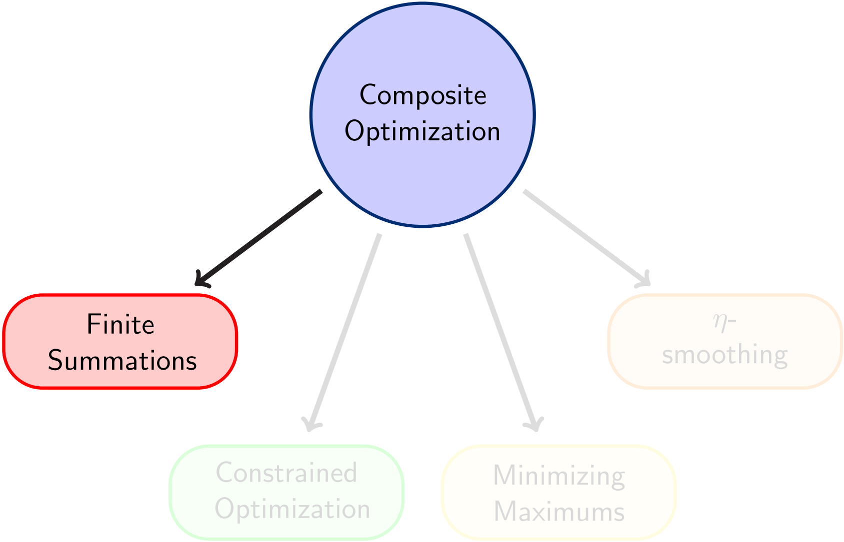


Compositions and Heterogeneity

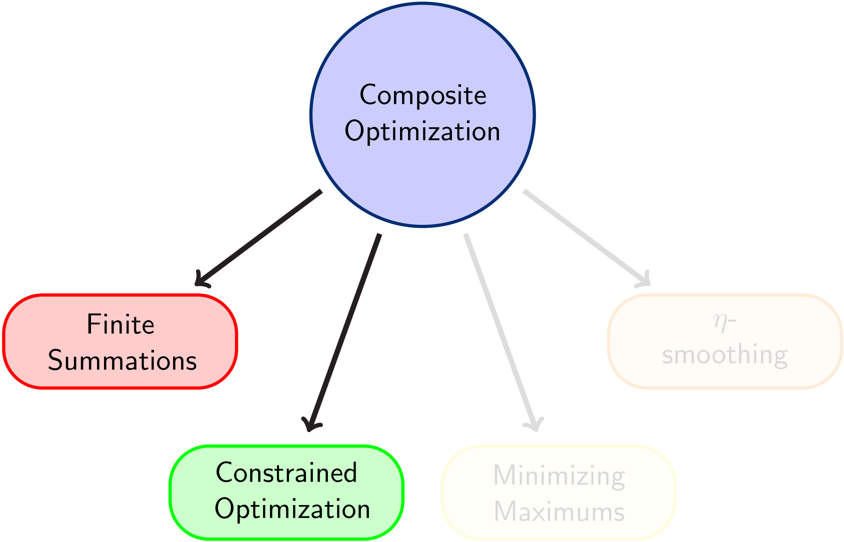
Composite is an Amazing Model



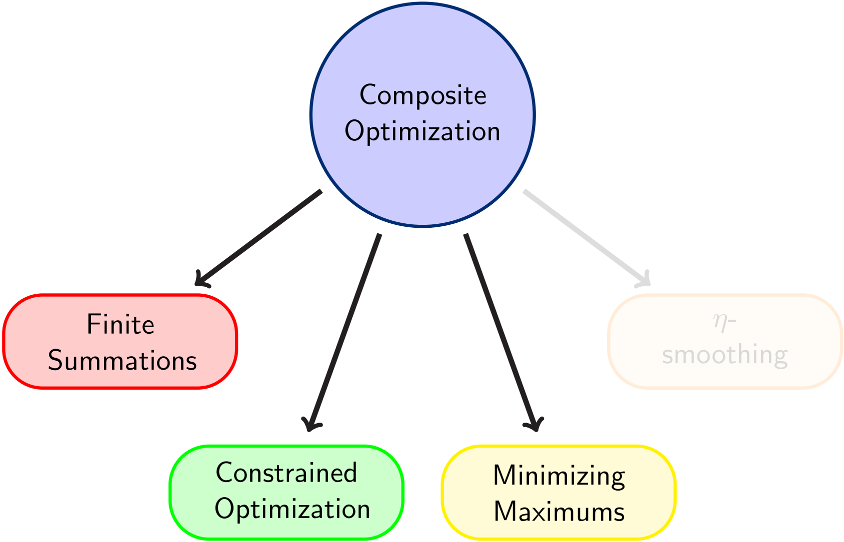
Composite is an Amazing Model



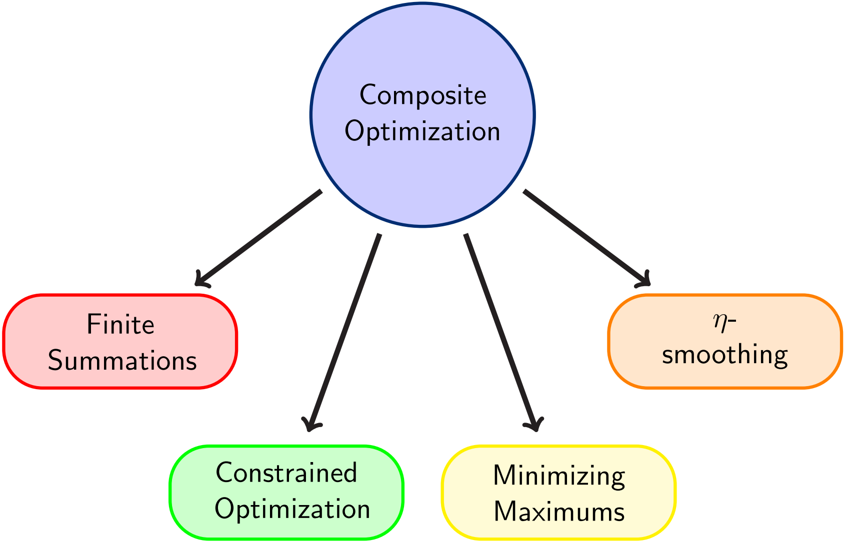
Composite is an Amazing Model



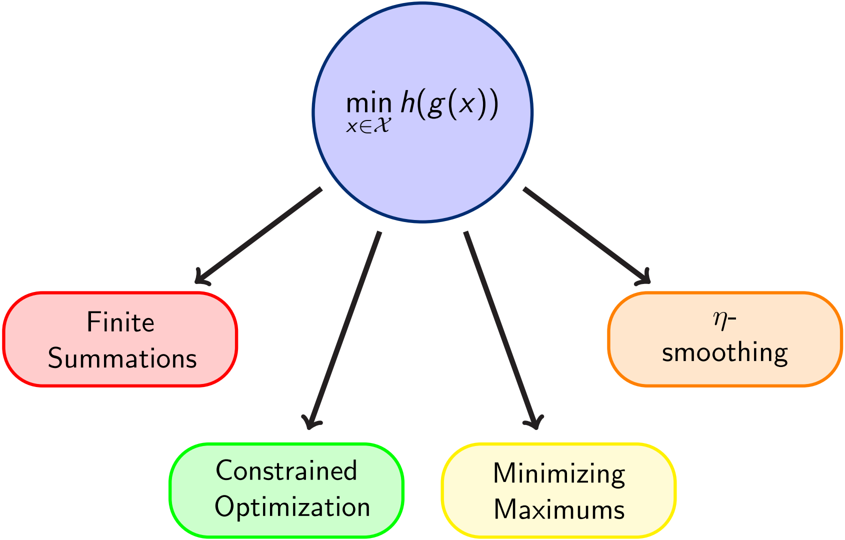
Composite is an Amazing Model



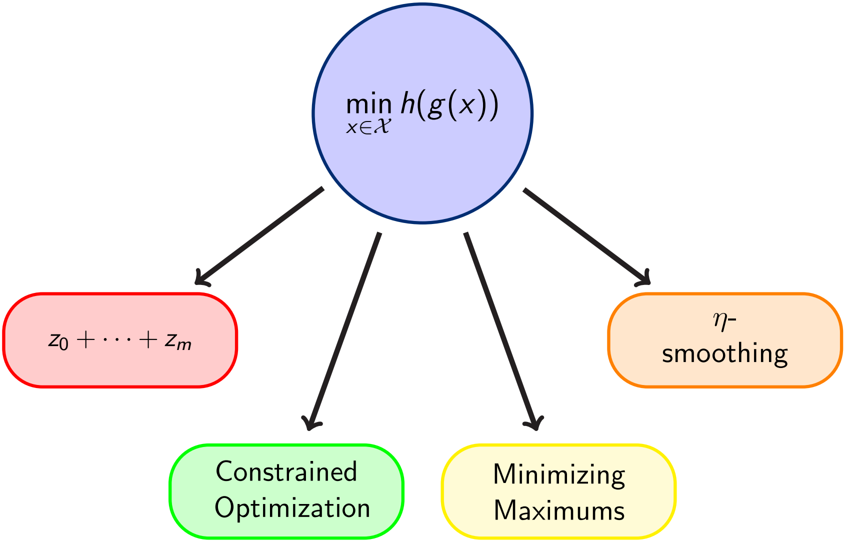
Composite is an Amazing Model



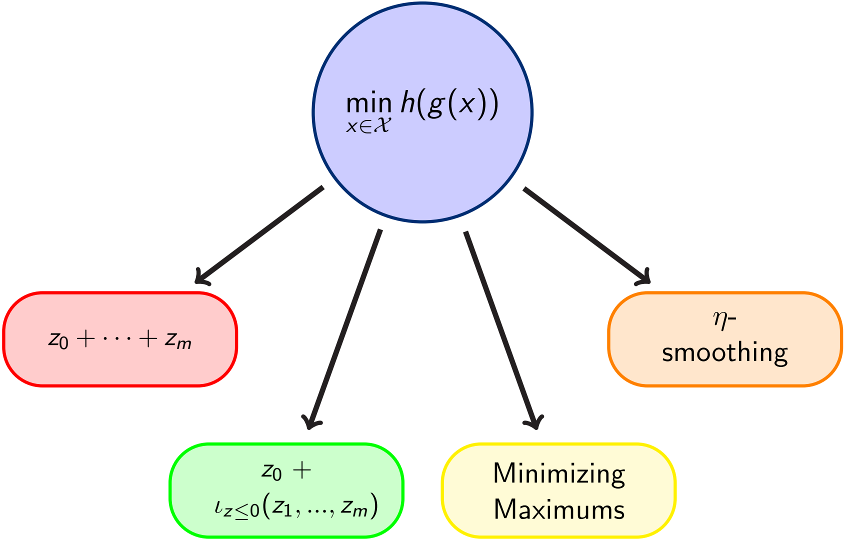
Composite is an Amazing Model



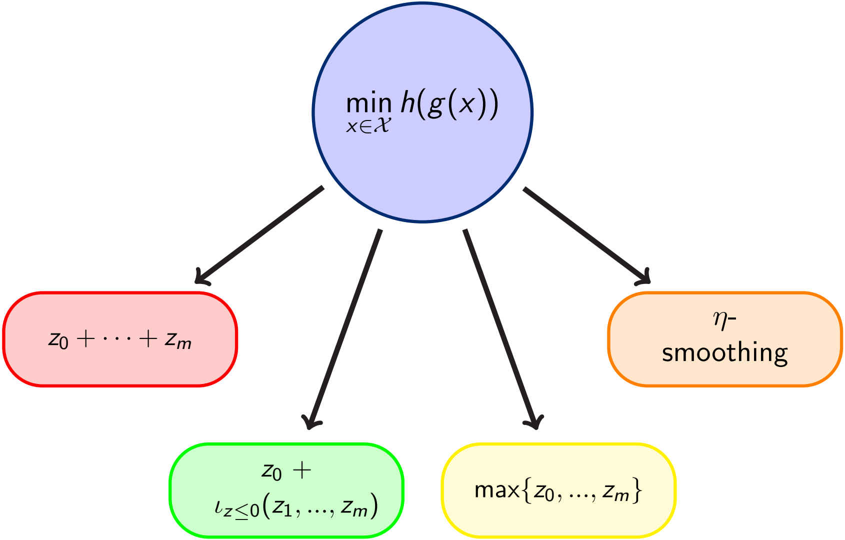
Composite is an Amazing Model



Composite is an Amazing Model



Composite is an Amazing Model



Composite is an Amazing Model

$$\min_{x\in\mathcal{X}} h(g(x))$$

$$z_0 + \cdots + z_m$$

$$z_0 + \iota_{z\le 0}(z_1, \ldots, z_m)$$

$$\max\{z_0, \ldots, z_m\}$$

$$\frac{1}{\eta}\log\sum_{j=0}^{m} e^{\eta z_j}$$

Heterogeneity is an Amazing Model

Consider the following ML problem of **support vector machines**

$$\begin{cases} \min\limits_{w,b,\xi} & \|w\|_2^2 + C\sum_{i=1}^{n} \xi_i \\ s.t. & y_i(w^T x_i - b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{cases}$$

$$\min_{w,b} \|w\|_2^2 + C\sum_{i=1}^{n} \max\{0, 1 - y_i(w^T x_i - b)\}$$

Heterogeneity is an Amazing Model

Consider the following ML problem of **support vector machines**

$$\begin{cases} \min\limits_{w,b,\xi} & \|w\|_2^2 + C\sum\limits_{i=1}^{n}\xi_i \\ s.t. & y_i(w^T x_i - b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{cases}$$

$$\min_{w,b} \|w\|_2^2 + C\sum_{i=1}^{n}\max\{0, 1 - y_i(w^T x_i - b)\}$$

Heterogeneity is an Amazing Model

We may have some adversarial type problem

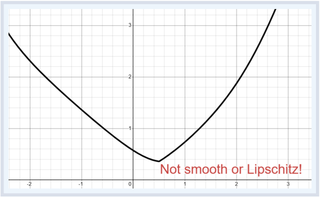
$$\min_{x \in \mathcal{X}} \max_{p \in \{1, \frac{4}{3}, \frac{3}{2}, 2\}} \left\{ \|Ax - b\|_p^p \right\}$$

Heterogeneity is an Amazing Model

Perhaps we're analyzing a mixture model or log-likelihood

$$\min_{x \in \mathcal{X}} \log\left(\sum_{j=1}^{m} \exp(g_j(x))\right)$$

$$\begin{aligned} g_1(x) &= x^2 \\ g_2(x) &= |2x - 1| \end{aligned}$$



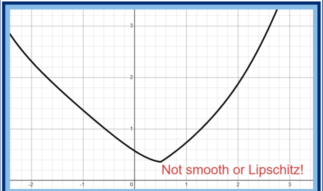
Note: Structured components $\implies$ structured objective!

Heterogeneity is an Amazing Model

Perhaps we're analyzing a mixture model or log-likelihood

$$\min_{x \in \mathcal{X}} \log\left(\sum_{j=1}^{m} \exp(g_j(x))\right)$$

$$\begin{aligned} g_1(x) &= x^2 \\ g_2(x) &= |2x-1| \end{aligned}$$



Note: Structured components $\Longrightarrow$ structured objective!

Smoothness and Convexity (Generalized)

Two Dual Notions

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

"L-smoothness"

"μ-strong convexity"

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

Two Dual Notions

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

"L-smoothness"

"μ-strong convexity"

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

Two Dual Notions

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$$

"L-smoothness"

"μ-strong convexity"

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2$$

Smooth and Strong Convexity

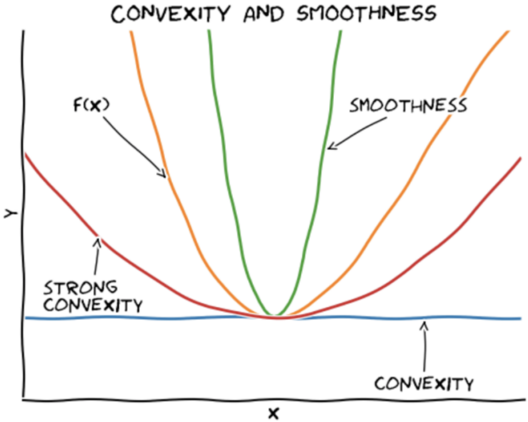
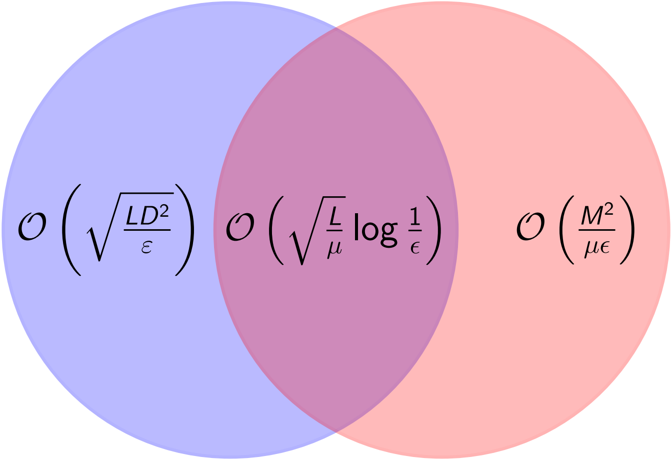


Figure: [2] Smoothness and Strong Convexity Visualized

Smooth and Strong Convexity Rates

L-**smooth** μ-**Strongly convex**[1]



[1]In the nonsmooth case, we assume f is Lipschitz with rank M

Generalizing Smoothness

Smoothness seems pretty restrictive.

Can we generalize, still using first order information?

Generalizing Smoothness

Smoothness seems pretty restrictive.

Can we generalize, still using first order information?

(L, p)-Hölder Smoothness

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|^p$$

(L, p)-Hölder continuous gradient

$(p + 1)$-degree upper bound

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{1+p}\|y - x\|^{1+p}$$

(L, p)-Hölder Smoothness

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|^p$$

(L, p)-Hölder continuous gradient

$(p + 1)$-degree upper bound

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{1+p}\|y - x\|^{1+p}$$

(μ, q)-Uniform Convexity

Can we analogously generalize the strong convexity?

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{1+q} \|y - x\|^{1+q}$$

(μ, q)-Uniform Convexity

Can we analogously generalize the strong convexity?

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{1+q} \|y - x\|^{1+q}$$

Visual Interlude

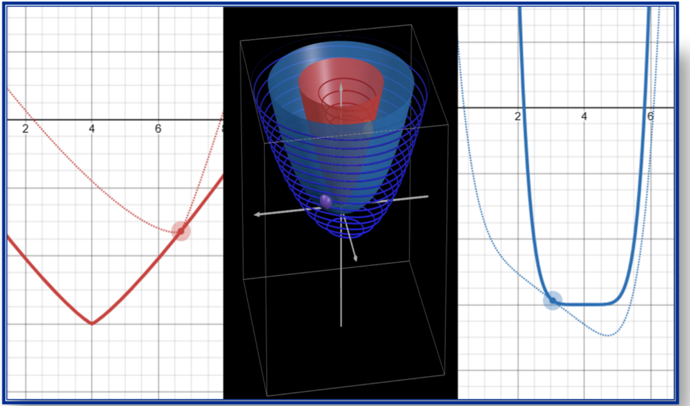


Figure: 2-dim plot and 3-dim plot

Universal Guarantees?

Recall, we minimize $F(x) = h(g_1(x), \ldots, g_m(x))$. Each component having its own (L_j, p_j)-Hölder smoothness and (μ_j, q_j)-uniform convexity.

Suppose we're given magic $\tilde{L}$ and $\tilde{\mu}$ that captures all the information for upper/lower curvature...

What guarantees should we hope for?

Universal Guarantees?

Recall, we minimize $F(x) = h(g_1(x), \ldots, g_m(x))$. Each component having its own (L_j, p_j)-Hölder smoothness and (μ_j, q_j)-uniform convexity.

Suppose we're given magic $\tilde{L}$ and $\tilde{\mu}$ that captures all the information for upper/lower curvature...

What guarantees should we hope for?

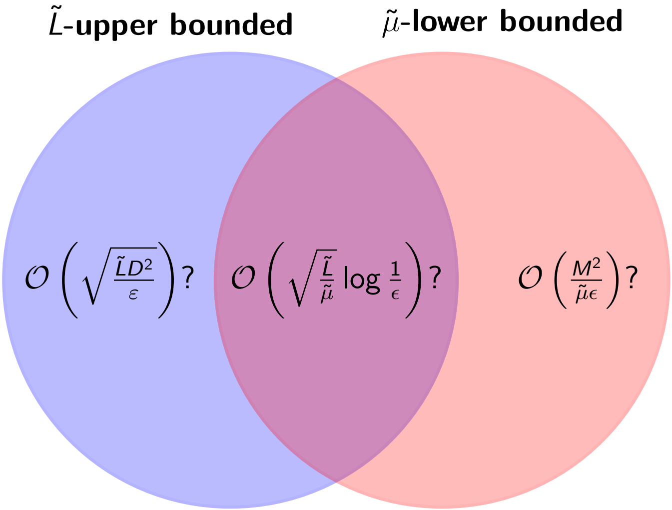
Universal Guarantees?

Recall, we minimize $F(x) = h(g_1(x), \ldots, g_m(x))$. Each component having its own (L_j, p_j)-Hölder smoothness and (μ_j, q_j)-uniform convexity.

Suppose we're given magic $\tilde{L}$ and $\tilde{\mu}$ that captures all the information for upper/lower curvature...

What guarantees should we hope for?

Aggregated Smooth and Strong Convexity Rates?



Motivating our Lovely Constants

From Constraints to Compositions

Constrained Optimization $\longrightarrow$ $$\min_{x\in\mathcal{X}} \; g_0(x) \quad \text{s.t.} \; g_j(x) \le 0$$

Composite Optimization $\longrightarrow$ $$\min_{x\in\mathcal{X}} \; g_0(x) + h(g_1(x), \dots, g_m(x))$$

From Constraints to Compositions

Constrained Optimization $\longrightarrow$ $\min_{x \in \mathcal{X}} g_0(x) + \iota_{z \le 0}(g_1(x), \ldots, g_m(x))$

Composite Optimization $\longrightarrow$ $\min_{x \in \mathcal{X}} g_0(x) + h(g_1(x), \ldots, g_m(x))$

From Constraints to Compositions

Constrained Optimization $\longrightarrow$ $$\min_{x\in\mathcal{X}} \; g_0(x) + \iota_{z\leq 0}(g_1(x),\dots,g_m(x))$$

Composite Optimization $\longrightarrow$ $$\min_{x\in\mathcal{X}} \; g_0(x) + h(g_1(x),\dots,g_m(x))$$

De "composing"

$$p_\star = \begin{cases} \min_{x\in\mathcal{X}} & g_0(x) \\ \text{s.t.} & g_j(x) \le 0 \end{cases}$$

$$p_\star = \min_{x\in\mathcal{X}} \max_{\lambda_j \ge 0} g_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x)$$

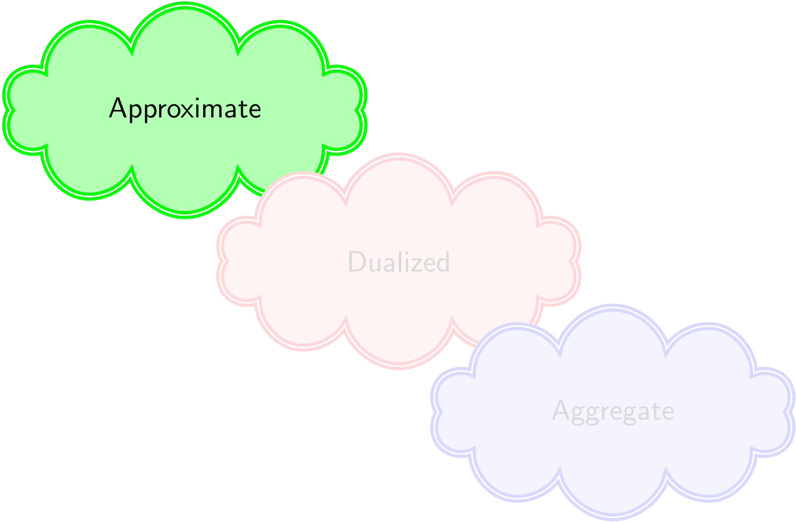
De "composing"

$$p_\star = \begin{cases} \min\limits_{x \in \mathcal{X}} & g_0(x) \\ \text{s.t.} & g_j(x) \leq 0 \end{cases}$$

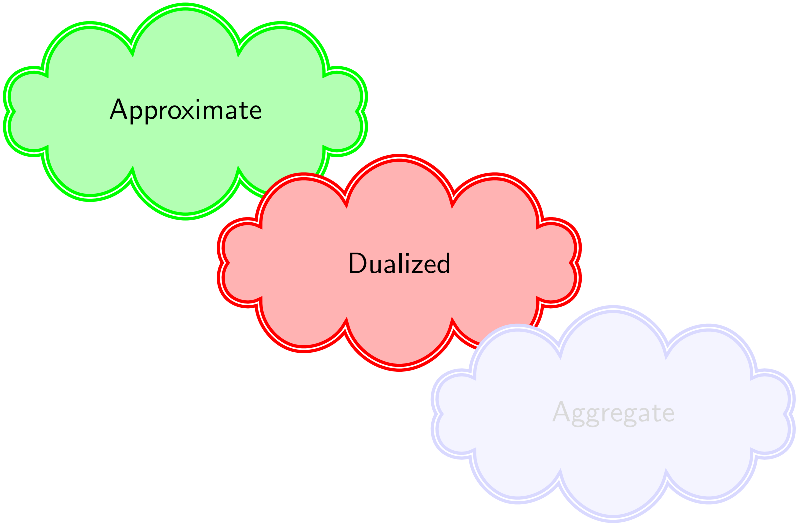
$$p_\star = \min_{x \in \mathcal{X}} \max_{\lambda_j \geq 0} g_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x)$$

Three Notions

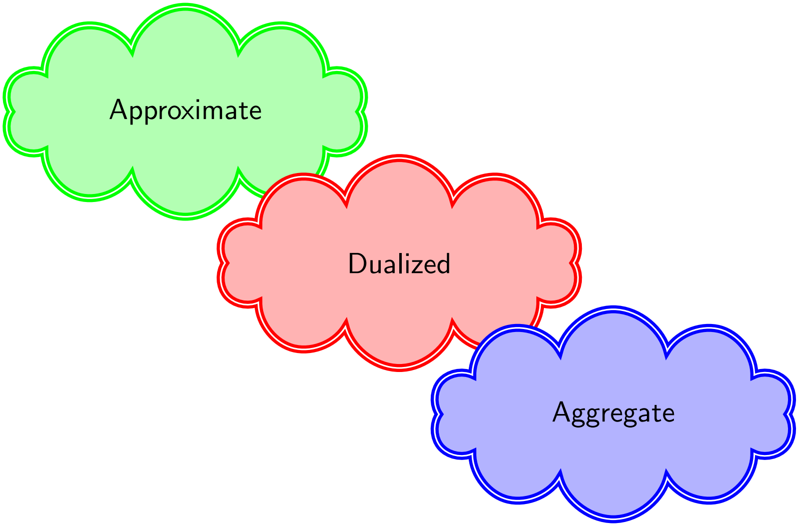
Three Notions



Three Notions



Three Notions



Tool 1: Fenchel Conjugates

Dualized

$$f^*(\lambda) = \sup_{x \in \mathbb{R}^n} \langle \lambda, x \rangle - f(x)$$

$$\Big\downarrow\; f^{**} = f$$

$$f(x) = \sup_{\lambda \in \mathbb{R}^n} \langle x, \lambda \rangle - f^*(\lambda)$$

Tool 1: Fenchel Conjugates

Dualized

$$f^*(\lambda) = \sup_{x \in \mathbb{R}^n} \langle \lambda, x \rangle - f(x)$$

$$\Big\downarrow \; f^{**} = f$$

$$f(x) = \sup_{\lambda \in \mathbb{R}^n} \langle x, \lambda \rangle - f^*(\lambda)$$

Tool 1: Fenchel Conjugates

Dualized

$$h^*(\lambda) = \sup_{x \in \mathbb{R}^n} \langle \lambda, x \rangle - h(x)$$

$$h^{**} = h$$

$$h(g(x)) = \sup_{\lambda \in \mathbb{R}^n} \langle \lambda, g(x) \rangle - h^*(\lambda)$$

Lagrangian Reformulation (via conjugates)

$$p_\star = \min_{x\in\mathcal{X}} g_0(x) + \boxed{\iota_{z\le 0}(g_1(x), \dots, g_m(x))}$$

$$p_\star = \min_{x\in\mathcal{X}} g_0(x) + \boxed{\max_{\lambda\in\mathbb{R}^m} \sum_{j=1}^{m} \lambda_j g_j(x) - \iota^*_{z\le 0}(\lambda)}$$

Lagrangian Reformulation (via conjugates)

$$p_\star = \min_{x\in\mathcal{X}} g_0(x) + \boxed{\iota_{z\leq 0}(g_1(x), \ldots, g_m(x))}$$

$$p_\star = \min_{x\in\mathcal{X}} g_0(x) + \boxed{\max_{\lambda\in\mathbb{R}^m} \sum_{j=1}^m \lambda_j g_j(x) - \iota^*_{z\leq 0}(\lambda)}$$

Lagrangian Reformulation (via conjugates)

$$p_\star = \min_{x\in\mathcal{X}} g_0(x) + \boxed{\iota_{z\leq 0}(g_1(x),\dots,g_m(x))}$$

$$p_\star = \min_{x\in\mathcal{X}} \max_{\lambda_j\geq 0} g_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x)$$

De "composing" (again, generally)

$$p_\star = \min_{x\in\mathcal{X}} g_0(x) + \boxed{h(g_1(x), \ldots, g_m(x))}$$

$$h(g(x)) = \max_{\lambda\in\mathbb{R}^m} \sum_{j=1}^m \lambda_j g_j(x) - h^*(\lambda)$$

$$p_\star = \min_{x\in\mathcal{X}} \max_{\lambda\geq 0} g_0(x) + \sum_{j=1}^m \lambda_j g_j(x) - h^*(\lambda)$$

De“composing” (again, generally)

$$p_\star = \min_{x\in\mathcal{X}} g_0(x) + \boxed{h(g_1(x), \ldots, g_m(x))}$$

$$h(g(x)) = \max_{\lambda\in\mathbb{R}^m} \sum_{j=1}^{m} \lambda_j g_j(x) - h^*(\lambda)$$

$$p_\star = \min_{x\in\mathcal{X}} \max_{\lambda\geq 0} g_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x) - h^*(\lambda)$$

Sums are Wonderful

Aggregate

$$\min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda^\star) := g_0(x) + \sum_{j=1}^{m} \lambda_j^\star g_j(x)$$

each g_j is L_j-smooth

$$\sum_{j=0}^{m} \lambda_j^\star g_j(x) \text{ is } \sum_{j=0}^{m} \lambda_j^\star L_j\text{-smooth}$$

Sums are Wonderful

Aggregate

$$\min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda^\star) := g_0(x) + \sum_{j=1}^{m} \lambda_j^\star g_j(x)$$

each g_j is L_j-smooth

$$\sum_{j=0}^{m} \lambda_j^\star g_j(x) \text{ is } \sum_{j=0}^{m} \lambda_j^\star L_j\text{-smooth}$$

Inaccessible $\lambda^\star$!

Approximate

$$\min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda) := g_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x)$$

restrict $\lambda \in \Lambda_r := B(\lambda^\star, r)$

$$\sum_{j=0}^{m} \lambda_j g_j(x) \text{ is } \sum_{j=0}^{m} (\lambda_j^\star + r) L_j\text{-smooth}$$

Inaccessible $\lambda^\star$!

Approximate

$$\min_{x \in \mathcal{X}} \mathcal{L}(x, \lambda) := g_0(x) + \sum_{j=1}^{m} \lambda_j g_j(x)$$

$$\Big\downarrow \text{ restrict } \lambda \in \Lambda_r := B(\lambda^\star, r)$$

$$\sum_{j=0}^{m} \lambda_j g_j(x) \text{ is } \sum_{j=0}^{m} (\lambda_j^\star + r) L_j\text{-smooth}$$

Defining Our Constants

Approximate Dualized Aggregate Smoothness Constant I

When g_j is L_j-smooth

WLOG let $g_0(x) = 0$

$$\boxed{L^{\mathtt{ADA}}_{\varepsilon,r} := \sum_{j=1}^{m} (\lambda_j^\star + r) L_j}$$

Tool 2: Nesterov-Style Quadratic Upper Bounds

Lemma (Lemma 1, Nesterov [1])

Fix $\delta > 0$ and (L, p)-Hölder smooth, with $L_\delta \geq \left[\frac{1-p}{1+p}\frac{1}{\delta}\right]^{\frac{1-p}{1+p}} L^{\frac{2}{1+p}}$,

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L_\delta}{2}\|y - x\|^2 + \frac{\delta}{2}, \ \forall x, y \in dom\ f \ . \tag{1}$$

Approximate Dualized Aggregate Smoothness Constant II

When g_j is (L_j, p_j)-Hölder smooth

setting $\delta = \frac{\varepsilon}{\sqrt{16 L^{\mathrm{ADA}}_{\varepsilon,r} D_x^2/\varepsilon}}$

$$L_{\delta,r} := \sum_{j=1}^{m} \left[\left[\frac{1-p_j}{1+p_j} \cdot \frac{m}{\delta} \right]^{\frac{1-p_j}{1+p_j}} \left[(\lambda_j^\star + r) \cdot L_j \right]^{\frac{2}{1+p_j}} \right]$$

Approximate Dualized Aggregate Smoothness Constant II

When g_j is (L_j, p_j)-Hölder smooth

setting $\delta = \dfrac{\varepsilon}{\sqrt{16 L^{\mathtt{ADA}}_{\varepsilon,r} D_x^2/\varepsilon}}$

$$L^{\mathtt{ADA}}_{\varepsilon,r} := \left\{ L^{\mathtt{ADA}} > 0 : L^{\mathtt{ADA}} = \sum_{j=1}^{m} \left[\frac{1-p_j}{1+p_j} \cdot \frac{m\sqrt{L^{\mathtt{ADA}}}}{\varepsilon} \cdot \frac{4D_x}{\sqrt{\varepsilon}} \right]^{\frac{1-p_j}{1+p_j}} \left[(\lambda_j^\star + r) L_j \right]^{\frac{2}{1+p_j}} \right\}$$

Tool 3: Restarting Methods

Suppose
we're given

Guarantees that
$f(x^N) - f(x^\star) \leq \frac{L\|x^0 - x^\star\|^2}{2N^2}$

Our algorithm produces output
s.t. $f(x^k) - f(x^\star) \leq \varepsilon$

$f(x^k) - f(x^\star) \geq \frac{\mu}{2}\|x^k - x^\star\|^2$

Tool 3: Restarting Methods

Suppose we're given

Guarantees that
$$f(x^N) - f(x^\star) \leq \frac{L\|x^0 - x^\star\|^2}{2N^2}$$

Our algorithm produces output s.t. $f(x^k) - f(x^\star) \leq \varepsilon$

$$f(x^k) - f(x^\star) \geq \frac{\mu}{2}\|x^k - x^\star\|^2$$

Tool 3: Restarting Methods

Suppose we're given

Guarantees that
$f(x^N) - f(x^\star) \leq \frac{L\|x^0 - x^\star\|^2}{2N^2}$

Our algorithm produces output
s.t. $f(x^k) - f(x^\star) \leq \varepsilon$

$f(x^k) - f(x^\star) \geq \frac{\mu}{2}\|x^k - x^\star\|^2$

Tool 3: Restarting Methods

Suppose we're given

Guarantees that $f(x^N) - f(x^\star) \leq \frac{L\|x^0 - x^\star\|^2}{2N^2}$

Our algorithm produces output s.t. $f(x^k) - f(x^\star) \leq \varepsilon$

$$f(x^k) - f(x^\star) \geq \frac{\mu}{2}\|x^k - x^\star\|^2$$

Tool 3: Restarting Methods

How many iterations to reach $\varepsilon/2$?

Recall we have
$$\|x^k - x^\star\|^2 \le \frac{2\varepsilon}{\mu}$$

After restarting
$$f(x^N) - f(x^\star) \le \frac{L\|x^k - x^\star\|^2}{2N^2}$$

$N = \sqrt{2L/\mu}$ yields
$$f(x^N) - f(x^\star) \le \frac{\varepsilon}{2}$$

Tool 3: Restarting Methods

Recall we have
$\|x^k - x^\star\|^2 \leq \frac{2\varepsilon}{\mu}$

How many iterations to reach $\varepsilon/2$?

After restarting
$f(x^N) - f(x^\star) \leq \frac{L\|x^k - x^\star\|^2}{2N^2}$

$N = \sqrt{2L/\mu}$ yields
$f(x^N) - f(x^\star) \leq \frac{\varepsilon}{2}$

Tool 3: Restarting Methods

How many iterations to reach $\varepsilon/2$?

Recall we have
$$\|x^k - x^\star\|^2 \;\leq\; \frac{2\varepsilon}{\mu}$$

After restarting
$$f(x^N) - f(x^\star) \leq \frac{L\|x^k - x^\star\|^2}{2N^2}$$

$N = \sqrt{2L/\mu}$ yields
$$f(x^N) - f(x^\star) \;\leq\; \frac{\varepsilon}{2}$$

Tool 3: Restarting Methods

How many iterations to reach $\varepsilon/2$?

Recall we have
$$\|x^k - x^\star\|^2 \ \leq \ \frac{2\varepsilon}{\mu}$$

After restarting
$$f(x^N) - f(x^\star) \leq \frac{L\|x^k - x^\star\|^2}{2N^2}$$

$N \ = \ \sqrt{2L/\mu}$ yields
$$f(x^N) - f(x^\star) \ \leq \ \frac{\varepsilon}{2}$$

General Growth Condition

Suppose $f(x) = \sum_{j=1}^{m} \lambda_j^\star g_j(x)$

Each g_j is (μ_j, q_j)-uniformly convex

$$f(x) - f(x^\star) \geq \sum_{j=1}^{m} \lambda_j^\star \frac{\mu_j}{q_j+1} \|x - x^\star\|^{q_j+1}$$

General Growth Condition

Suppose $f(x) = \sum_{j=1}^{m} \lambda_j^\star g_j(x)$

Each g_j is (μ_j, q_j)-uniformly convex

$$f(x) - f(x^\star) \geq \sum_{j=1}^{m} \lambda_j^\star \frac{\mu_j}{q_j + 1} \|x - x^\star\|^{q_j+1}$$

General Growth Condition

Suppose $f(x) \;=\; \sum_{j=1}^{m} \lambda_j^\star g_j(x)$

$$\Big\Downarrow \text{ Each } g_j \text{ is } (\mu_j, q_j)\text{-uniformly convex}$$

$$f(x) - f(x^\star) \geq \underbrace{\sum_{j=1}^{m} \lambda_j^\star \frac{\mu_j}{q_j+1} \|x - x^\star\|^{q_j+1}}_{G_x(\|x - x^\star\|)}$$

Revisiting Restarting

Suppose we're given

Guarantees that
$f(x^N) - f(x^\star) \leq \frac{L^{\mathtt{ADA}}_{\varepsilon,r}\|x^0 - x^\star\|^2}{2N^2}$

Our algorithm produces output s.t. $f(x^k) - f(x^\star) \leq \varepsilon$

$f(x^k) - f(x^\star) \geq G_x\left(\|x^k - x^\star\|\right)$

Revisiting Restarting

Guarantees that

$$f(x^N) - f(x^\star) \le \frac{L^{\mathtt{ADA}}_{\varepsilon,r}\|x^0 - x^\star\|^2}{2N^2}$$

Suppose we're given

Our algorithm produces output s.t. $f(x^k) - f(x^\star) \le \varepsilon$

$$f(x^k) - f(x^\star) \ge G_x\left(\|x^k - x^\star\|\right)$$

Revisiting Restarting

Guarantees that

$$f(x^N) - f(x^\star) \leq \frac{L^{\mathtt{ADA}}_{\varepsilon,r}\|x^0 - x^\star\|^2}{2N^2}$$

Suppose we're given

Our algorithm produces output s.t. $f(x^k) - f(x^\star) \leq \varepsilon$

$$f(x^k) - f(x^\star) \geq G_x\left(\|x^k - x^\star\|\right)$$

Revisiting Restarting

Suppose we're given

Our algorithm produces output s.t. $f(x^k) - f(x^\star) \leq \varepsilon$

Guarantees that $f(x^N) - f(x^\star) \leq \frac{L^{\texttt{ADA}}_{\varepsilon,r}\|x^0 - x^\star\|^2}{2N^2}$

$$f(x^k) - f(x^\star) \geq G_x\left(\|x^k - x^\star\|\right)$$

Revisiting Restarting

How many iterations to reach $\varepsilon/2$?

Recall we have $\|x^k - x^\star\|^2 \leq (G_x^{-1}(\varepsilon))^2$

After restarting $f(x^N) - f(x^\star) \leq \frac{L^{\mathrm{ADA}}_{\varepsilon,r}\|x^k - x^\star\|^2}{2N^2}$

$N = \sqrt{\frac{2L^{\mathrm{ADA}}_{\varepsilon,r}(G_x^{-1}(\varepsilon))^2}{2\varepsilon}}$ yields $f(x^N) - f(x^\star) \leq \frac{\varepsilon}{2}$

Revisiting Restarting

Recall we have
$$\|x^k - x^\star\|^2 \le (G_x^{-1}(\varepsilon))^2$$

How many iterations to reach $\varepsilon/2$?

After restarting
$$f(x^N) - f(x^\star) \le \frac{L^{\mathrm{ADA}}_{\varepsilon,r}\|x^k - x^\star\|^2}{2N^2}$$

$$N = \sqrt{\frac{2L^{\mathrm{ADA}}_{\varepsilon,r}(G_x^{-1}(\varepsilon))^2}{2\varepsilon}}$$
yields $f(x^N) - f(x^\star) \le \frac{\varepsilon}{2}$

Revisiting Restarting

Recall we have
$$\|x^k - x^\star\|^2 \leq (G_x^{-1}(\varepsilon))^2$$

How many iterations to reach $\varepsilon/2$?

After restarting
$$f(x^N) - f(x^\star) \leq \frac{L^{\mathrm{ADA}}_{\varepsilon,r}\|x^k - x^\star\|^2}{2N^2}$$

$$N = \sqrt{\frac{2L^{\mathrm{ADA}}_{\varepsilon,r}(G_x^{-1}(\varepsilon))^2}{2\varepsilon}}$$
yields $f(x^N) - f(x^\star) \leq \frac{\varepsilon}{2}$

Revisiting Restarting

How many iterations to reach $\varepsilon/2$?

Recall we have
$$\|x^k - x^\star\|^2 \leq (G_x^{-1}(\varepsilon))^2$$

After restarting
$$f(x^N) - f(x^\star) \leq \frac{L^{\mathtt{ADA}}_{\varepsilon,r}\|x^k - x^\star\|^2}{2N^2}$$

$$N = \sqrt{\frac{2L^{\mathtt{ADA}}_{\varepsilon,r}(G_x^{-1}(\varepsilon))^2}{2\varepsilon}}$$
yields $f(x^N) - f(x^\star) \leq \frac{\varepsilon}{2}$

Revisiting Restarting

How many iterations to reach $\varepsilon/2$?

Recall we have
$$\|x^k - x^\star\|^2 \leq (G_x^{-1}(\varepsilon))^2$$

After restarting
$$f(x^N) - f(x^\star) \leq \frac{L^{\mathtt{ADA}}_{\varepsilon,r}\|x^k - x^\star\|^2}{2N^2}$$

$$N = \sqrt{\frac{2L^{\mathtt{ADA}}_{\varepsilon,r}(G_x^{-1}(\varepsilon))^2}{2\varepsilon}}$$
yields $f(x^N) - f(x^\star) \leq \frac{\varepsilon}{2}$

$\mu^{\mathtt{ADA}}_{\varepsilon}$?

Revisiting Restarting

How many iterations to reach $\varepsilon/2$?

Recall we have
$$\|x^k - x^\star\|^2 \leq (G_x^{-1}(\varepsilon))^2$$

After restarting
$$f(x^N) - f(x^\star) \leq \frac{L^{\mathtt{ADA}}_{\varepsilon,r}\|x^k - x^\star\|^2}{2N^2}$$

$$N = \sqrt{\frac{2L^{\mathtt{ADA}}_{\varepsilon,r}(G_x^{-1}(\varepsilon))^2}{2\varepsilon}}$$
yields $f(x^N) - f(x^\star) \leq \frac{\varepsilon}{2}$

$1/\mu^{\mathtt{ADA}}_{2\varepsilon}$

Approximate Dualized Aggregate Convexity Constant

When g_j is (μ_j, q_j)-uniformly convex

$$\mu_\varepsilon^{\texttt{ADA}} = \frac{\varepsilon}{(G_x^{-1}(\varepsilon/2))^2}$$

Approximate Dualized Aggregate Convexity Constant

When g_j is (μ_j, q_j)-uniformly convex

$$\mu_\varepsilon^{\mathtt{ADA}} := \left\{ \mu^{\mathtt{ADA}} > 0 : \frac{\mu^{\mathtt{ADA}}}{2} = \sum_{j=1}^{m} \lambda_j^\star \frac{\mu_j}{q_j+1} (\varepsilon/\mu^{\mathtt{ADA}})^{\frac{q_j-1}{2}} \right\}$$

Approximate Dualized Aggregate Smoothness Constant III

When g_j is (L_j, p_j)-Hölder smooth and (μ_j, q_j)-uniformly convex

$$L^{\text{ADA}}_{\varepsilon,r} := \left\{ L^{\text{ADA}} > 0 : L^{\text{ADA}} = \sum_{j=1}^{m} \left[\frac{1-p_j}{1+p_j} \cdot \frac{m\sqrt{L^{\text{ADA}}}}{\varepsilon} \cdot \min\left\{ \frac{4D_x}{\sqrt{\varepsilon}}, \frac{8}{\sqrt{\mu^{\text{ADA}}_{\varepsilon}}} \right\} \right]^{\frac{1-p_j}{1+p_j}} \left[(\lambda_j^\star + r) L_j \right]^{\frac{2}{1+p_j}} \right\}$$

Algorithms and Analysis

Extended Lagrangian and Gap Function

$$F(x) = h(\underbrace{g_1(x), \ldots, g_m(x)}_{g(x)}) + u(x)$$

$$\mathcal{L}(x; \lambda) := \langle \lambda, g(x) \rangle - h^*(\lambda) + u(x)$$

Extended Lagrangian and Gap Function

$$F(x) = h(\underbrace{g_1(x), \ldots, g_m(x)}_{g(x)}) + u(x)$$

$$\mathcal{L}(x;\lambda) := \langle \lambda, g(x)\rangle - h^*(\lambda) + u(x)$$

Extended Lagrangian and Gap Function

$$\mathcal{L}(x;\lambda) := \langle \lambda, g(x) \rangle - h^*(\lambda) + u(x)$$

$$\mathcal{L}(x;\lambda,\nu) := \langle \lambda, \underbrace{\nu x - g^*(\nu)}_{\text{conjugate of g}} \rangle - h^*(\lambda) + u(x)$$

Extended Lagrangian and Gap Function

$$\mathcal{L}(x;\lambda,\nu) := \langle \lambda, \underbrace{\nu x - g^*(\nu)}_{\text{conjugate of g}} \rangle - h^*(\lambda) + u(x)$$

Gap Function

$$Q(z,\hat{z}) = \mathcal{L}(x;\hat{\lambda},\hat{\nu}) - \mathcal{L}(\hat{x};\lambda,\nu)$$

Minimizing the Gap Function ("Q-Analysis" and Sliding)

$$Q(z^t; z) = Q_\nu(z^t; z) + Q_\lambda(z^t; z) + Q_x(z^t; z)$$

$$\begin{aligned} &\mathcal{L}(x^t; \lambda, \nu) - \mathcal{L}(x^t; \lambda, \nu^t) \\ = \quad & \langle \lambda, \nu x^t - g^*(\nu) \rangle \boxed{-\langle \lambda, \nu^t x^t - g^*(\nu^t) \rangle} \end{aligned}$$

$$\begin{aligned} &\mathcal{L}(x^t; \lambda, \nu^t) - \mathcal{L}(x^t; \lambda^t, \nu^t) \\ = \; & \langle \lambda, \nu^t x^t - g^*(\nu^t) \rangle - h^*(\lambda) \boxed{- [\langle \lambda^t, \nu^t x^t - g^*(\nu^t) \rangle - h^*(\lambda^t)]} \end{aligned}$$

$$\begin{aligned} &\mathcal{L}(x^t; \lambda^t, \nu^t) - \mathcal{L}(x; \lambda^t, \nu^t) \\ = \quad & \boxed{\langle \sum_{i=1}^m \lambda_i^t \nu_i^t, x^t \rangle + u(x^t)} - \langle \textstyle\sum_{i=1}^m \lambda_i^t \nu_i^t, x \rangle - f(x) \end{aligned}$$

Minimizing the Gap Function ("Q-Analysis" and Sliding)

$$Q(z^t; z) = Q_\nu(z^t; z) + Q_\lambda(z^t; z) + Q_x(z^t; z)$$

$$\mathcal{L}(x^t; \lambda, \nu) - \mathcal{L}(x^t; \lambda, \nu^t)$$
$$= \langle \lambda, \nu x^t - g^*(\nu)\rangle \boxed{-\langle \lambda, \nu^t x^t - g^*(\nu^t)\rangle}$$

$$\mathcal{L}(x^t; \lambda, \nu^t) - \mathcal{L}(x^t; \lambda^t, \nu^t)$$
$$= \langle \lambda, \nu^t x^t - g^*(\nu^t)\rangle - h^*(\lambda) \boxed{-\left[\langle \lambda^t, \nu^t x^t - g^*(\nu^t)\rangle - h^*(\lambda^t)\right]}$$

$$\mathcal{L}(x^t; \lambda^t, \nu^t) - \mathcal{L}(x; \lambda^t, \nu^t)$$
$$= \boxed{\langle \sum_{i=1}^{m} \lambda_i^t \nu_i^t, x^t\rangle + u(x^t)} - \langle \textstyle\sum_{i=1}^m \lambda_i^t \nu_i^t, x\rangle - f(x)$$

Minimizing the Gap Function ("Q-Analysis" and Sliding)

$$Q(z^t; z) \quad = \quad Q_\nu(z^t; z) + Q_\lambda(z^t; z) + Q_x(z^t; z)$$

$$\begin{aligned} & \mathcal{L}(x^t; \lambda, \nu) - \mathcal{L}(x^t; \lambda, \nu^t) \\ = \quad & \langle \lambda, \nu x^t - g^*(\nu) \rangle \boxed{- \langle \lambda, \nu^t x^t - g^*(\nu^t) \rangle} \end{aligned}$$

$$\begin{aligned} & \mathcal{L}(x^t; \lambda, \nu^t) - \mathcal{L}(x^t; \lambda^t, \nu^t) \\ = \; & \langle \lambda, \nu^t x^t - g^*(\nu^t) \rangle - h^*(\lambda) \boxed{- \left[\langle \lambda^t, \nu^t x^t - g^*(\nu^t) \rangle - h^*(\lambda^t) \right]} \end{aligned}$$

$$\begin{aligned} & \mathcal{L}(x^t; \lambda^t, \nu^t) - \mathcal{L}(x; \lambda^t, \nu^t) \\ = \quad & \boxed{\langle \sum_{i=1}^m \lambda_i^t \nu_i^t, x^t \rangle + u(x^t)} - \langle \textstyle\sum_{i=1}^m \lambda_i^t \nu_i^t, x \rangle - f(x) \end{aligned}$$

Minimizing the Gap Function ("Q-Analysis" and Sliding)

$$Q(z^t; z) \quad = \quad Q_\nu(z^t; z) + Q_\lambda(z^t; z) + Q_x(z^t; z)$$

$$\begin{aligned} &\mathcal{L}(x^t; \lambda, \nu) - \mathcal{L}(x^t; \lambda, \nu^t) \\ = \quad &\langle \lambda, \nu x^t - g^*(\nu) \rangle \boxed{-\langle \lambda, \nu^t x^t - g^*(\nu^t) \rangle} \end{aligned}$$

$$\begin{aligned} &\mathcal{L}(x^t; \lambda, \nu^t) - \mathcal{L}(x^t; \lambda^t, \nu^t) \\ = \; &\langle \lambda, \nu^t x^t - g^*(\nu^t) \rangle - h^*(\lambda) \boxed{- \left[\langle \lambda^t, \nu^t x^t - g^*(\nu^t) \rangle - h^*(\lambda^t) \right]} \end{aligned}$$

$$\begin{aligned} &\mathcal{L}(x^t; \lambda^t, \nu^t) - \mathcal{L}(x; \lambda^t, \nu^t) \\ = \quad &\boxed{\langle \sum_{i=1}^m \lambda_i^t \nu_i^t, x^t \rangle + u(x^t)} - \langle \textstyle\sum_{i=1}^m \lambda_i^t \nu_i^t, x \rangle - f(x) \end{aligned}$$

Minimizing the Gap Function ("Q-Analysis" and Sliding)

$$Q(z^t; z) \quad = \quad Q_\nu(z^t; z) \; + \; Q_\lambda(z^t; z) \; + \; Q_x(z^t; z)$$

$$\nu_j^t \;\; \leftarrow \text{argmax}_{\nu_j \in V_j} \langle \nu, \tilde{x}^t \rangle \; - \; g_j^*(\nu) \; - \; \tau_t U_{g_j^*}(\nu_j; \nu_j^{t-1})$$

$$\begin{aligned} & \mathcal{L}(x^t; \lambda, \nu^t) - \mathcal{L}(x^t; \lambda^t, \nu^t) \\ = \; & \langle \lambda, \nu^t x^t - g^*(\nu^t) \rangle \; - \; h^*(\lambda) \boxed{- \left[\langle \lambda^t, \nu^t x^t - g^*(\nu^t) \rangle - h^*(\lambda^t) \right]} \end{aligned}$$

$$\begin{aligned} & \mathcal{L}(x^t; \lambda^t, \nu^t) - \mathcal{L}(x; \lambda^t, \nu^t) \\ = \; & \boxed{\langle \sum_{i=1}^m \lambda_i^t \nu_i^t, x^t \rangle + u(x^t)} \; - \; \langle \textstyle\sum_{i=1}^m \lambda_i^t \nu_i^t, x \rangle \; - \; f(x) \end{aligned}$$

Minimizing the Gap Function ("Q-Analysis" and Sliding)

$$Q(z^t; z) \quad = \quad Q_\nu(z^t; z) \, + \, Q_\lambda(z^t; z) \, + \, Q_x(z^t; z)$$

$$\nu_j^t \;\leftarrow \text{argmax}_{\nu_j \in V_j} \langle \nu, \tilde{x}^t \rangle \; - \; g_j^*(\nu) \; - \; \tau_t U_{g_j^*}(\nu_j; \nu_j^{t-1})$$

$$\lambda^t \leftarrow \text{argmax}_{\lambda \in \Lambda} \langle \lambda, \nu^t \tilde{x}^t - g^*(\nu^t) \rangle - h^*(\lambda) - \tfrac{\gamma_t}{2} \|\lambda - \lambda^{t-1}\|^2$$

$$\begin{aligned} & \mathcal{L}(x^t; \lambda^t, \nu^t) - \mathcal{L}(x; \lambda^t, \nu^t) \\ = \; & \boxed{\langle \sum_{i=1}^{m} \lambda_i^t \nu_i^t, x^t \rangle + u(x^t)} \; - \; \langle \textstyle\sum_{i=1}^{m} \lambda_i^t \nu_i^t, x \rangle \; - \; f(x) \end{aligned}$$

Minimizing the Gap Function ("Q-Analysis" and Sliding)

$$Q(z^t; z) \;=\; Q_\nu(z^t; z) + Q_\lambda(z^t; z) + Q_x(z^t; z)$$

$$\nu_j^t \leftarrow \mathrm{argmax}_{\nu_j \in V_j} \langle \nu, \tilde{x}^t \rangle - g_j^*(\nu) - \tau_t U_{g_j^*}(\nu_j; \nu_j^{t-1})$$

$$\lambda^t \leftarrow \mathrm{argmax}_{\lambda \in \Lambda} \langle \lambda, \nu^t \tilde{x}^t - g^*(\nu^t) \rangle - h^*(\lambda) - \tfrac{\gamma_t}{2} \|\lambda - \lambda^{t-1}\|^2$$

$$x^t \leftarrow \mathrm{argmin}_{x \in \mathcal{X}} \langle \textstyle\sum_{j=1}^m \lambda_j^t \nu_j^t, x \rangle + u(x) + \tfrac{\eta_t}{2} \|x - x^{t-1}\|^2$$

Minimizing the Gap Function ("Q-Analysis" and Sliding)

$$Q(z^t; z) \quad = \quad Q_\nu(z^t; z) + Q_\lambda(z^t; z) + Q_x(z^t; z)$$

$$\nu^t \leftarrow \nabla g(\underline{x}^t),\ \underline{x}^t \leftarrow \tfrac{\tau_t \underline{x}^{t-1} + \tilde{x}^t}{1+\tau_t} \text{ with } \tilde{x}^t = x^{t-1} + \theta_t(x^{t-1} - x^{t-2})$$

$$\lambda^t \leftarrow \text{argmax}_{\lambda \in \Lambda} \langle \lambda, \nu^t \tilde{x}^t - g^*(\nu^t) \rangle - h^*(\lambda) - \tfrac{\gamma_t}{2} \|\lambda - \lambda^{t-1}\|^2$$

$$x^t \leftarrow \text{argmin}_{x \in \mathcal{X}} \langle \textstyle\sum_{j=1}^m \lambda_j^t \nu_j^t, x \rangle + u(x) + \tfrac{\eta_t}{2} \|x - x^{t-1}\|^2$$

The Universal Fast Composite Method (UFCM)

Algorithm 1 Universal Fast Composite Method (UFCM)

Input $z^0 \in \mathcal{X} \times \Lambda$, outer loop iteration count T, and smoothness constant $L^{\mathsf{ADA}}_{\varepsilon,r}$

Initialize $x^{-1} = \underline{x}^0 = y_0^{(1)} = x^0 \in \mathcal{X}$, $\lambda_{-1}^{(1)} = \lambda_0^{(1)} = \lambda^0 \in \Lambda$, and parameters $\{\theta_t\}, \{\eta_t\}, \{\tau_t\}, \{\omega_t\}$ as a function of $L^{\mathsf{ADA}}_{\varepsilon,r}$

Set $\nu^0 = \nabla g(x^0)$.
for $t = 1,\ 2,\ 3,\ ...,\ T$ **do**
  Set $\underline{x}^t \leftarrow (\tau_t \underline{x}^{t-1} + \tilde{x}^t)/(1+\tau_t)$ where $\tilde{x}^t = x^{t-1} + \theta_t(x^{t-1} - x^{t-2})$
  Set $\nu^t \leftarrow \nabla g(\underline{x}^t)$
  Calculate inner loop iteration limit S_t, paramters $\beta^{(t)}$, $\gamma^{(t)}$, and $\rho^{(t)}$
  for $s = 1, 2, ..., S_t$ **do**
    Set $\tilde{h}^{(t),s} = \begin{cases} (\nu^t)^T \lambda_0^{(t)} + \rho^{(t)} (\nu^{t-1})^T (\lambda_0^{(t)} - \lambda_{-1}^{(t)}) & \text{if } s = 1, \\ (\nu^t)^T \lambda_{s-1}^{(t)} + (\nu^t)^T (\lambda_{s-1}^{(t)} - \lambda_{s-2}^{(t)}) & \text{otherwise} \end{cases}$
    Solve $y_s^{(t)} \leftarrow \operatorname*{argmin}_{y \in \mathcal{X}} \left\langle \tilde{h}^{(t),s}, y \right\rangle + u(y) + \frac{\eta_t}{2}\|y - x^{t-1}\|^2 + \frac{\beta^{(t)}}{2}\|y - y_{s-1}^{(t)}\|^2$
    Solve $\lambda_s^{(t)} \leftarrow \operatorname*{argmax}_{\lambda \in \Lambda} \left\langle \lambda, \nu^t (y_s^{(t)} - \underline{x}^t) + g(\underline{x}^t) \right\rangle - h^*(\lambda) - \frac{\gamma^{(t)}}{2}\|\lambda - \lambda_{s-1}^{(t)}\|^2$
  end for
  Set $\lambda_0^{(t+1)} = \lambda_{S_t}^{(t)}$, $\lambda_{-1}^{(t+1)} = \lambda_{S_t-1}^{(t)}$, $y_0^{(t+1)} = y_{S_t}^{(t)}$
  Set $x^t = \sum_{s=1}^{S_t} y_s^{(t)}/S_t$ and $\tilde{\lambda}^t = \sum_{s=1}^{S_t} \lambda_s^{(t)}/S_t$
end for
return $(\bar{x}^T, \bar{\lambda}^T) := \sum_{t=1}^{T} \omega_t \left(x^t, \tilde{\lambda}^t\right) / \left(\sum_{t=1}^{T} \omega_t\right)$

Figure: Modified from [3]

Restarted-UFCM

Algorithm 2 Restarted Universal Fast Composite Method (R-UFCM)

Input $z^0 \in \mathcal{X} \times \Lambda$, distance bounds D_x and D_λ, target accuracy $\varepsilon > 0$, constants $L^{\mathsf{ADA}}_{\varepsilon,r}$ and $\mu^{\mathsf{ADA}}_{\varepsilon}$, and UFCM execution count $K = \left\lceil \log_2\left(\frac{Q(\bar{z}^0,\hat{z})+\varepsilon}{\varepsilon}\right)\right\rceil$

Set $D_x^{(0)}$, $D_\lambda^{(0)}$ and $\{T_k\}$ according to (5.3)

for $k = 0, 1, \ldots, K-1$ **do**

  Run UFCM(z^k, $\lceil T_k \rceil$, $L^{\mathsf{ADA}}_{\varepsilon,r}$) returning output $(\bar{x}^{T_k,k}, \bar{\lambda}^{T_k,k})$

  Set $(x^{k+1}, D_x^{(k+1)}) = \begin{cases} (\bar{x}^{T_k,k}, \sqrt{2^{K-k}\varepsilon/\mu^{\mathsf{ADA}}_{\varepsilon}}) & \text{if } \mu^{\mathsf{ADA}}_{\varepsilon} \geq 4\varepsilon/D_x^2 \\ (x^0, D_x) & \text{otherwise} \end{cases}$

  Set $(\lambda^{k+1}, D_\lambda^{(k+1)}) = \begin{cases} (\bar{\lambda}^{T_k,k}, \sqrt{2^{K-k}\varepsilon L_h}) & \text{if } L_h \leq D_\lambda^2/\varepsilon \\ (\lambda^0, D_\lambda) & \text{otherwise} \end{cases}$

  Set $z^{k+1} = (x^{k+1}, \lambda^{k+1})$

end for

Figure: Restarted Variant

Universally Optimal Guarantees

Theorem 1: $\mathcal{O}\left(\sqrt{\frac{L^{\text{ADA}}_{\varepsilon,r} D_x^2}{\mu^{\text{ADA}}_{\varepsilon}}}\right)$

total gradient calls to g

when the objective

is sufficiently convex

$(\mu^{\text{ADA}}_{\varepsilon} \geq \varepsilon / D_x^2)$

Theorem 2: $\mathcal{O}\left(\sqrt{\frac{L^{\text{ADA}}_{\varepsilon,r}}{\mu^{\text{ADA}}_{\varepsilon}}} \log\left(\frac{1}{\varepsilon}\right)\right)$

total gradient calls to g

Universally Optimal Guarantees

Theorem 1: $\mathcal{O}\left(\sqrt{\frac{L^{\mathtt{ADA}}_{\varepsilon,r} D_x^2}{\mu^{\mathtt{ADA}}_{\varepsilon}}}\right)$ total gradient calls to g

when the objective is sufficiently convex $(\mu^{\mathtt{ADA}}_{\varepsilon} \geq \varepsilon/D_x^2)$

Theorem 2: $\mathcal{O}\left(\sqrt{\frac{L^{\mathtt{ADA}}_{\varepsilon,r}}{\mu^{\mathtt{ADA}}_{\varepsilon}}} \log\left(\frac{1}{\varepsilon}\right)\right)$ total gradient calls to g

Immediate Corollaries

Suppose g is (L, p)-Hölder smooth

$$\mathcal{O}\left(\left(\frac{L}{\varepsilon}\right)^{\frac{2}{1+3p}}\|x^0 - x^\star\|^{\frac{2+2p}{1+3p}}\right)$$

total gradient calls to g

Immediate Corollaries

Suppose g is (L, p)-Hölder smooth

$$\mathcal{O}\left(\left(\frac{L}{\varepsilon}\right)^{\frac{2}{1+3p}} \left\|x^0 - x^\star\right\|^{\frac{2+2p}{1+3p}}\right)$$

total gradient calls to g

Immediate Corollaries

$$L^{\texttt{ADA}}_{\varepsilon,r} = (1+r)^{\frac{4}{1+3p}} \left[\frac{1-p}{1+p} \cdot \frac{4D_x}{\varepsilon\sqrt{\varepsilon}}\right]^{\frac{2-2p}{1+3p}} L^{\frac{4}{1+3p}}$$

$$\begin{aligned} &\mathcal{O}\left(\sqrt{\frac{L^{\texttt{ADA}}_{\varepsilon,r} D_x^2}{\varepsilon}}\right) \\ &= \mathcal{O}\left(\sqrt{\left(\frac{D_x}{\varepsilon\sqrt{\varepsilon}}\right)^{\frac{2-2p}{1+3p}} \frac{L^{\frac{4}{1+3p}} D_x^2}{\varepsilon}}\right) \\ &= \mathcal{O}\left(\left(\frac{L}{\varepsilon}\right)^{\frac{2}{1+3p}} D_x^{\frac{2+2p}{1+3p}}\right) \end{aligned}$$

Immediate Corollaries

$$L_{\varepsilon,r}^{\mathtt{ADA}} = (1+r)^{\frac{4}{1+3p}} \left[\frac{1-p}{1+p} \cdot \frac{4D_x}{\varepsilon\sqrt{\varepsilon}}\right]^{\frac{2-2p}{1+3p}} L^{\frac{4}{1+3p}}$$

$$\begin{aligned} &\mathcal{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\mathtt{ADA}} D_x^2}{\varepsilon}}\right) \\ &= \mathcal{O}\left(\sqrt{\left(\frac{D_x}{\varepsilon\sqrt{\varepsilon}}\right)^{\frac{2-2p}{1+3p}} \frac{L^{\frac{4}{1+3p}} D_x^2}{\varepsilon}}\right) \\ &= \mathcal{O}\left(\left(\frac{L}{\varepsilon}\right)^{\frac{2}{1+3p}} D_x^{\frac{2+2p}{1+3p}}\right) \end{aligned}$$

Immediate Corollaries II

Suppose g is (L,p)-Hölder smooth *and* (μ,q)-uniformly convex

$$\begin{cases} \mathcal{O}\left(\left(\frac{L^{1+q}}{\mu^{1+p}\varepsilon^{q-p}}\right)^{\frac{2}{(1+3p)(1+q)}}\right) & \text{if } q>p\ , \\ \mathcal{O}\left(\left(\frac{L^{1+q}}{\mu^{1+p}}\right)^{\frac{2}{(1+q)(1+3p)}}\log\left(\frac{G(x^0)-G^\star}{\varepsilon}\right)\right) & \text{if } q=p \end{cases}$$

total gradient calls to g

Immediate Corollaries II

Suppose g is (L, p)-Hölder smooth *and* (μ, q)-uniformly convex

$$\begin{cases} \mathcal{O}\left(\left(\frac{L^{1+q}}{\mu^{1+p}\varepsilon^{q-p}}\right)^{\frac{2}{(1+3p)(1+q)}}\right) & \text{if } q > p\ , \\ \mathcal{O}\left(\left(\frac{L^{1+q}}{\mu^{1+p}}\right)^{\frac{2}{(1+q)(1+3p)}} \log\left(\frac{G(x^0)-G^\star}{\varepsilon}\right)\right) & \text{if } q = p \end{cases}$$

total gradient calls to g

Immediate Corollaries II

$$L^{\mathtt{ADA}}_{\varepsilon,r} = \left[\frac{1-p}{1+p} \cdot \frac{8}{\varepsilon\sqrt{\mu^{\mathtt{ADA}}_{\varepsilon}}}\right]^{\frac{2-2p}{1+3p}} L^{\frac{4}{1+3p}}$$

$$\mu^{\mathtt{ADA}}_{\varepsilon} = 2\left(\frac{\mu}{1+q}\right)^{\frac{2}{1+q}} \varepsilon^{\frac{q-1}{q+1}}$$

$$\tilde{O}\left(\sqrt{\frac{L^{\mathtt{ADA}}_{\varepsilon,r}}{\mu^{\mathtt{ADA}}_{\varepsilon}}}\right) = \tilde{O}\left(\frac{\left(\varepsilon\sqrt{\mu^{\mathtt{ADA}}_{\varepsilon}}\right)^{-\frac{1-p}{1+3p}} L^{\frac{2}{1+3p}}}{\sqrt{\mu^{\mathtt{ADA}}_{\varepsilon}}}\right) = \tilde{O}\left(\left(\frac{L^{1+q}}{\mu^{1+p}\varepsilon^{q-p}}\right)^{\frac{2}{(1+3p)(1+q)}}\right),$$

Immediate Corollaries II

$$L_{\varepsilon,r}^{\texttt{ADA}} = \left[\frac{1-p}{1+p} \cdot \frac{8}{\varepsilon\sqrt{\mu_\varepsilon^{\texttt{ADA}}}}\right]^{\frac{2-2p}{1+3p}} L^{\frac{4}{1+3p}}$$

$$\mu_\varepsilon^{\texttt{ADA}} = 2\left(\frac{\mu}{1+q}\right)^{\frac{2}{1+q}} \varepsilon^{\frac{q-1}{q+1}}$$

$$\begin{aligned} &\tilde{O}\left(\sqrt{\frac{L_{\varepsilon,r}^{\texttt{ADA}}}{\mu_\varepsilon^{\texttt{ADA}}}}\right) \\ &= \tilde{O}\left(\frac{\left(\varepsilon\sqrt{\mu_\varepsilon^{\texttt{ADA}}}\right)^{-\frac{1-p}{1+3p}} L^{\frac{2}{1+3p}}}{\sqrt{\mu_\varepsilon^{\texttt{ADA}}}}\right) \\ &= \tilde{O}\left(\left(\frac{L^{1+q}}{\mu^{1+p}\varepsilon^{q-p}}\right)^{\frac{2}{(1+3p)(1+q)}}\right), \end{aligned}$$

References

Yurii Nesterov.
Universal gradient methods for convex optimization problems.
Mathematical Programming, 152(1–2):381–404, May 2014.

Sebastian Pokutta.
Cheat sheet: Smooth convex optimization.
https://www.pokutta.com/blog/research/2018/12/06/cheatsheet-smooth-idealized.html.

Zhe Zhang and Guanghui Lan.
Solving convex smooth function constrained optimization is almost as easy as unconstrained optimization.
arXiv preprint arXiv:2210.05807, 2022.