Principled approaches and tools for the analysis and design of first-order optimization algorithms

Adrien Taylor
Sierra team, Inria Paris

Optimization for Learning – Elliit symposium, May 2026

François Glineur

Julien Hendrickx

Aymeric Dieuleveut

Baptiste Goujaud

Céline Moucer

Yoel Drori

Francis Bach

Laurent Lessard

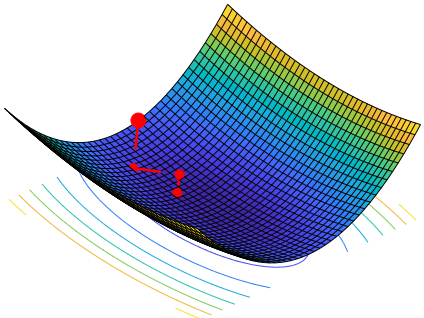
Bryan Van Scoy

Pontus Giselsson

Sebastian Banert

Manu Upadhyaya

Usually solved via **iterative algorithm** generating sequence $x_0, x_1, \ldots, x_N$.



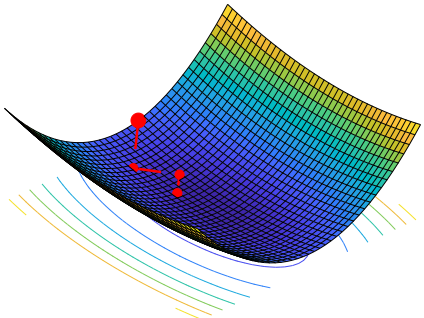
Gradient descent (stepsize α)

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

$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

end for

Usually solved via **iterative algorithm** generating sequence $x_0, x_1, \ldots, x_N$.



Gradient descent (stepsize α)

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

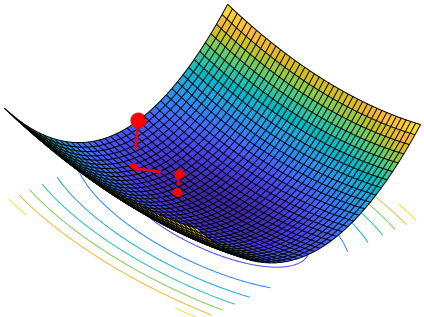
$$x_{k+1} = x_k - \alpha \nabla f(x_k)$$

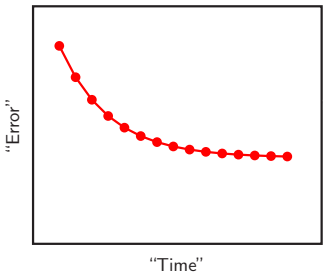
end for

What to expect from the output of the algorithm?

For instance: **bounds** on certain notions of "error": $f(x_k) - f(x_\star)$, $\|x_k - x_\star\|$, $\|\nabla f(x_k)\|$, etc.

Usually solved via **iterative algorithm** generating sequence $x_0, x_1, \ldots, x_N$.

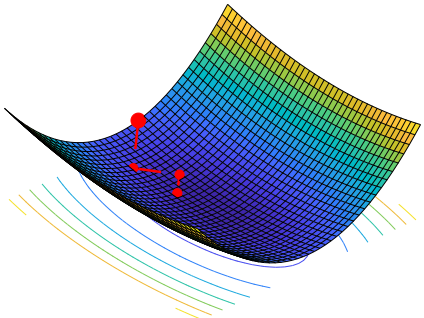


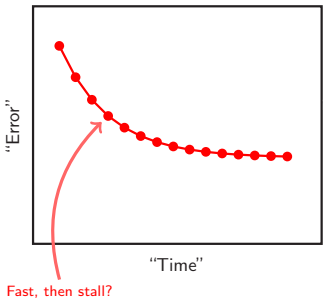


What to expect from the output of the algorithm?

For instance: **bounds** on certain notions of "error": $f(x_k) - f(x_\star)$, $\|x_k - x_\star\|$, $\|\nabla f(x_k)\|$, etc.

Usually solved via **iterative algorithm** generating sequence $x_0, x_1, \ldots, x_N$.

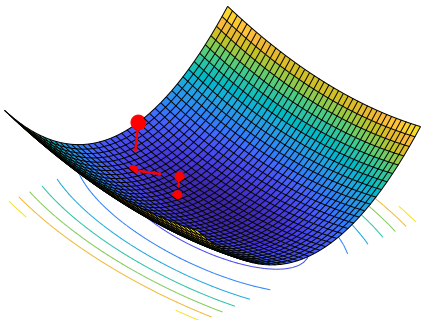


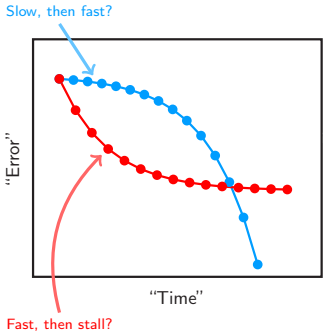


What to expect from the output of the algorithm?

For instance: **bounds** on certain notions of "error": $f(x_k) - f(x_\star)$, $\|x_k - x_\star\|$, $\|\nabla f(x_k)\|$, etc.

Usually solved via **iterative algorithm** generating sequence $x_0, x_1, \ldots, x_N$.

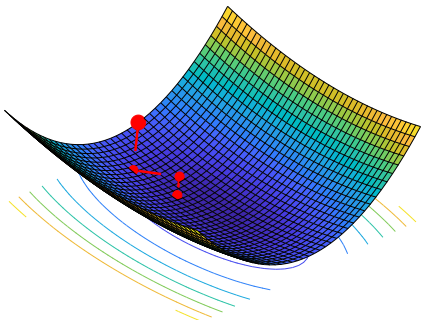


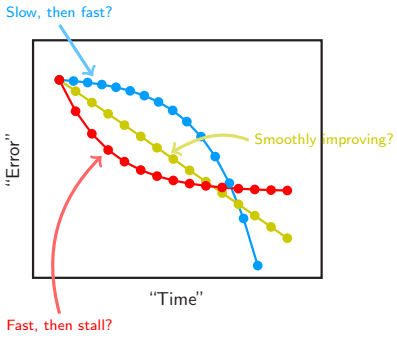


What to expect from the output of the algorithm?

For instance: **bounds** on certain notions of "error": $f(x_k) - f(x_\star)$, $\|x_k - x_\star\|$, $\|\nabla f(x_k)\|$, etc.

Usually solved via **iterative algorithm** generating sequence $x_0, x_1, \ldots, x_N$.

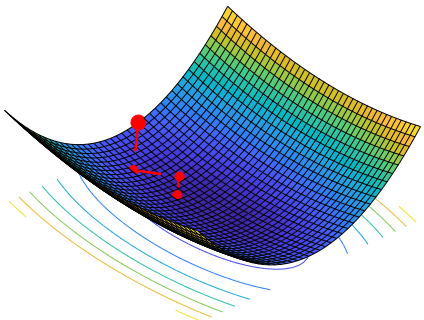


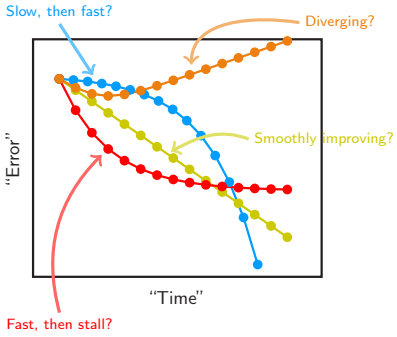


What to expect from the output of the algorithm?

For instance: **bounds** on certain notions of "error": $f(x_k) - f(x_\star)$, $\|x_k - x_\star\|$, $\|\nabla f(x_k)\|$, etc.

Usually solved via **iterative algorithm** generating sequence $x_0, x_1, \ldots, x_N$.





What to expect from the output of the algorithm?

For instance: **bounds** on certain notions of "error": $f(x_k) - f(x_\star)$, $\|x_k - x_\star\|$, $\|\nabla f(x_k)\|$, etc.

How to show that an algorithm works?

How to show that an algorithm works?

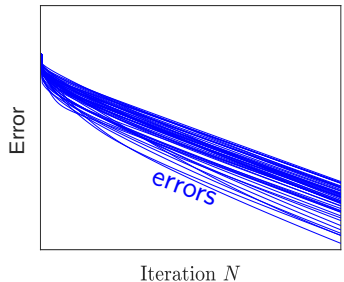
- Assumptions (no free lunch).

How to show that an algorithm works?

- ◇ Assumptions (no free lunch).
- ◇ Here: worst-case perspective.

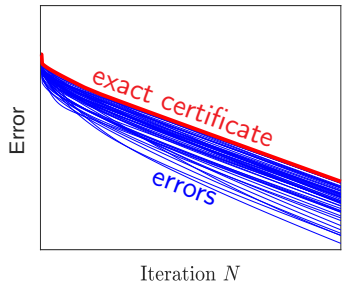
How to show that an algorithm works?

- ◇ Assumptions (no free lunch).
- ◇ Here: worst-case perspective.



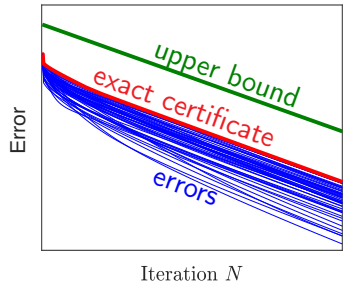
How to show that an algorithm works?

- ◇ Assumptions (no free lunch).
- ◇ Here: worst-case perspective.



How to show that an algorithm works?

- ◇ Assumptions (no free lunch).
- ◇ Here: worst-case perspective.



Constructive approach to performance analysis

Structured analyses

Towards optimal algorithms

Concluding remarks

Constructive approach to performance analysis

Structured analyses

Towards optimal algorithms

Concluding remarks

Example: analysis of a gradient method

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable). Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x),$$

with f is L-smooth and μ-strongly convex ($f \in \mathcal{F}_{\mu,L}$).

Example: analysis of a gradient method

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable). Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x),$$

with f is L-smooth and μ-strongly convex ($f \in \mathcal{F}_{\mu,L}$).

(Gradient method) We decide to use: $x_{k+1} = x_k - \alpha \nabla f(x_k)$

Example: analysis of a gradient method

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable). Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x),$$

with f is L-smooth and μ-strongly convex ($f \in \mathcal{F}_{\mu,L}$).

(Gradient method) We decide to use: $x_{k+1} = x_k - \alpha \nabla f(x_k)$

Question: what *a priori* guarantees after n iterations?

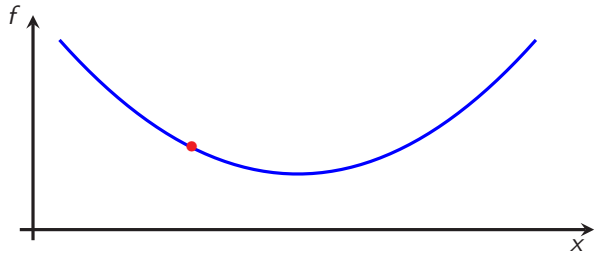
Examples: what about $f(x_n) - f(x_\star)$, $\|\nabla f(x_n)\|$, $\|x_n - x_\star\|$?

Step 1: collect inequalities

Differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:

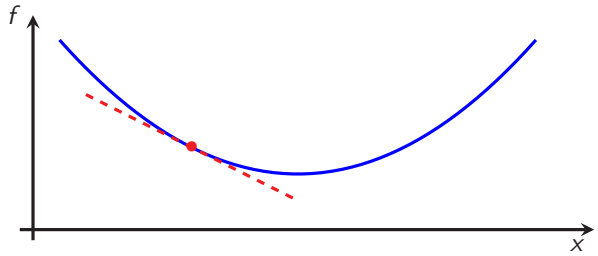
Step 1: collect inequalities

Differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:



Step 1: collect inequalities

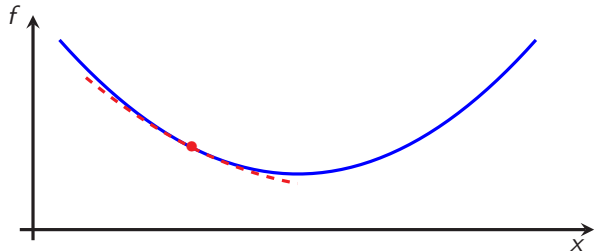
Differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:



(1) (Convexity) $f(x) \geqslant f(y) + \langle \nabla f(y), x - y \rangle$,

Step 1: collect inequalities

Differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:

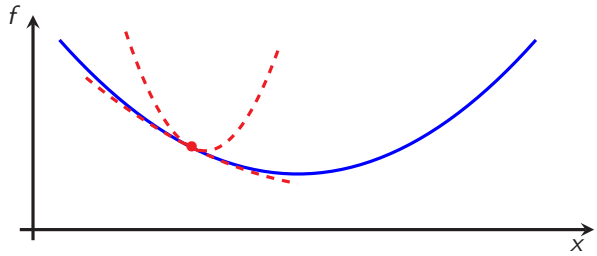


(1) (Convexity) $f(x) \geqslant f(y) + \langle \nabla f(y), x - y \rangle$,

(1b) (μ-strong convexity) $f(x) \geqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2$,

Step 1: collect inequalities

Differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:



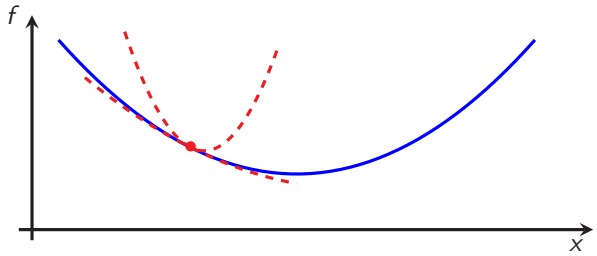
(1) (Convexity) $f(x) \geqslant f(y) + \langle \nabla f(y), x - y \rangle$,

(1b) (μ-strong convexity) $f(x) \geqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2$,

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

Step 1: collect inequalities

Differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex and L-smooth iff $\forall x, y \in \mathbb{R}^d$:



(1) (Convexity) $f(x) \geqslant f(y) + \langle \nabla f(y), x - y \rangle$,

(1b) (μ-strong convexity) $f(x) \geqslant f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2$,

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

(1&2) $\langle \nabla f(x) - \nabla f(y); x - y \rangle \geqslant \frac{1}{L+\mu}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{\mu L}{L+\mu}\|x - y\|^2$.

Step 2: combine inequalities

Step 2: combine inequalities

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

Step 2: combine inequalities

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),

Step 2: combine inequalities

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,

Step 2: combine inequalities

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

Step 2: combine inequalities

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

$$\|x_1 - x_\star\|^2$$

Step 2: combine inequalities

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

$$\|x_1 - x_\star\|^2 = \|x_0 - x_\star\|^2 - 2\alpha \langle \nabla f(x_0); x_0 - x_\star \rangle + \alpha^2 \|\nabla f(x_0)\|^2$$

Step 2: combine inequalities

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

$$\|x_1 - x_\star\|^2 = \|x_0 - x_\star\|^2 - 2\alpha \langle \nabla f(x_0); x_0 - x_\star \rangle + \alpha^2 \|\nabla f(x_0)\|^2$$

$$\Big\downarrow \text{ Inequality (1\&2)}$$

Step 2: combine inequalities

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

$$\|x_1 - x_\star\|^2 = \|x_0 - x_\star\|^2 - 2\alpha \langle \nabla f(x_0); x_0 - x_\star \rangle + \alpha^2 \|\nabla f(x_0)\|^2$$

$$\Big\downarrow \text{ Inequality (1\&2)}$$

$$\leqslant \left(1 - \tfrac{2\alpha L \mu}{L+\mu}\right) \|x_0 - x_\star\|^2 + \alpha \left(\alpha - \tfrac{2}{L+\mu}\right) \|\nabla f(x_0)\|^2$$

Step 2: combine inequalities

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

$$\|x_1 - x_\star\|^2 = \|x_0 - x_\star\|^2 - 2\alpha \langle \nabla f(x_0); x_0 - x_\star \rangle + \alpha^2 \|\nabla f(x_0)\|^2$$

$$\Big\downarrow \text{ Inequality (1\&2)}$$

$$\leqslant \left(1 - \tfrac{2\alpha L \mu}{L+\mu}\right) \|x_0 - x_\star\|^2 + \alpha \left(\alpha - \tfrac{2}{L+\mu}\right) \|\nabla f(x_0)\|^2$$

$$\Big\downarrow \text{ if } 0 \leqslant \alpha \leqslant \tfrac{2}{L+\mu}$$

Step 2: combine inequalities

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

$$\begin{aligned}
\|x_1 - x_\star\|^2 &= \|x_0 - x_\star\|^2 - 2\alpha \langle \nabla f(x_0); x_0 - x_\star \rangle + \alpha^2 \|\nabla f(x_0)\|^2 \\
&\Big\downarrow \text{ Inequality (1\&2)} \\
&\leqslant \left(1 - \tfrac{2\alpha L \mu}{L+\mu}\right) \|x_0 - x_\star\|^2 + \alpha \left(\alpha - \tfrac{2}{L+\mu}\right) \|\nabla f(x_0)\|^2 \\
&\Big\downarrow \text{ if } 0 \leqslant \alpha \leqslant \tfrac{2}{L+\mu} \\
&\leqslant (1 - \alpha\mu)^2 \|x_0 - x_\star\|^2.
\end{aligned}$$

Convergence rate of a gradient step

Legitimate questions:

Convergence rate of a gradient step

Legitimate questions:

- ◇ anything improvable? Realistic analyses?

Convergence rate of a gradient step

Legitimate questions:

- ◇ anything improvable? Realistic analyses?
- ◇ How to choose the right inequalities to combine?

Convergence rate of a gradient step

Legitimate questions:

- ◇ anything improvable? Realistic analyses?
- ◇ How to choose the right inequalities to combine?
- ◇ Why studying this specific quantity? Possible to adapt to other quantities?

Convergence rate of a gradient step

Legitimate questions:

- ◇ anything improvable? Realistic analyses?
- ◇ How to choose the right inequalities to combine?
- ◇ Why studying this specific quantity? Possible to adapt to other quantities?
- ◇ Unique way to arrive to the desired result?

Convergence rate of a gradient step

Legitimate questions:

- ◇ anything improvable? Realistic analyses?
- ◇ How to choose the right inequalities to combine?
- ◇ Why studying this specific quantity? Possible to adapt to other quantities?
- ◇ Unique way to arrive to the desired result?
- ◇ How likely are we to find such proofs in more complicated cases?

Legitimate questions about performance analyses?

Lemma 3. *Assume that the function is L-smooth and μ strongly-convex and satisfies the strong-growth condition in Equation 10. Then, using the updates in Equation 3-5 and setting the parameters according to Equations 17-18 if $\eta \leq \frac{1}{\rho L}$, then the following relation holds:*

$$b_{k+1}^2\eta\left[\mathbb{E}f(w_{k+1}) - f^*\right] \leq \frac{\alpha_0^2}{\rho\eta}\left[f(x_0) - f^*\right] + \frac{b_0^2}{2\rho\eta}\left\|x_0 - w^*\right\|^2 + \frac{\alpha^2\eta}{\rho}\sum_{i=0}^{k}\left[\gamma_i^2 b_{i+1}^2\right]$$

Proof.

Let $r_{k+1} = \|w_{k+1} - w^*\|$, then using equation 5

$$\begin{aligned} r_{k+1}^2 &= \|\beta_k v_k + (1-\beta_k)\zeta_k - w^* - \gamma_k\eta\nabla f(\zeta_k, z_k)\|^2 \\ r_{k+1}^2 &= \|\beta_k v_k + (1-\beta_k)\zeta_k - w^*\|^2 + \gamma_k^2\eta^2\left\|\nabla f(\zeta_k, z_k)\right\|^2 + 2\gamma_k\eta\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k, z_k)\rangle \end{aligned}$$

Taking expectation wrt to z_k,

$$\begin{aligned} \mathbb{E}[r_{k+1}^2] &= \mathbb{E}\|\beta_k v_k + (1-\beta_k)\zeta_k - w^*\|^2 + \gamma_k^2\eta^2\mathbb{E}\left\|\nabla f(\zeta_k, z_k)\right\|^2 + 2\gamma_k\eta\mathbb{E}[\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k, z_k)\rangle] \\ &\leq \|\beta_k v_k + (1-\beta_k)\zeta_k - w^*\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle + \gamma_k^2\eta^2\sigma^2 \\ &= \|\beta_k(v_k - w^*) + (1-\beta_k)(\zeta_k - w^*)\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle + \gamma_k^2\eta^2\sigma^2 \\ &\leq \beta_k\|v_k - w^*\|^2 + (1-\beta_k)\|\zeta_k - w^*\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle + \gamma_k^2\eta^2\sigma^2 \\ &\qquad \text{(By convexity of } \|\cdot\|^2\text{)} \\ &= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle + \gamma_k^2\eta^2\sigma^2 \\ &= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\left\langle\frac{\beta_k(1-\alpha_k)}{\alpha_k}(w_k - \zeta_k) + w^* - \zeta_k, \nabla f(\zeta_k)\right\rangle + \gamma_k^2\eta^2\sigma^2 \\ &\qquad \text{(From equation 4)} \\ &= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}\langle\nabla f(\zeta_k), (w_k - \zeta_k)\rangle + \left\langle\nabla f(\zeta_k), w^* - \zeta_k\right\rangle\right] + \gamma_k^2\eta^2\sigma^2 \\ &\leq \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}(f(w_k) - f(\zeta_k)) + \langle\nabla f(\zeta_k), w^* - \zeta_k\rangle\right] + \gamma_k^2\eta^2\sigma^2 \\ &\qquad \text{(By convexity)} \end{aligned}$$

By strong convexity,

$$\begin{aligned} \mathbb{E}[r_{k+1}^2] &\leq \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 \\ &\quad + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}(f(w_k) - f(\zeta_k)) + f^* - f(\zeta_k) - \frac{\mu}{2}\left\|\zeta_k - w^*\right\|^2\right] + \gamma_k^2\eta^2\sigma^2 \end{aligned} \tag{16}$$

By Lipschitz continuity of the gradient,

$$\begin{aligned} f(w_{k+1}) - f(\zeta_k) &\leq \langle\nabla f(\zeta_k), w_{k+1} - \zeta_k\rangle + \frac{L}{2}\left\|w_{k+1} - \zeta_k\right\|^2 \\ &\leq -\eta\langle\nabla f(\zeta_k), \nabla f(\zeta_k, z_k)\rangle + \frac{L\eta^2}{2}\left\|\nabla f(\zeta_k, z_k)\right\|^2 \end{aligned}$$

Taking expectation wrt z_k and using equations 9, 10

$$\begin{aligned} \mathbb{E}\left[f(w_{k+1}) - f(\zeta_k)\right] &\leq -\eta\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\rho\eta^2}{2}\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2} \\ \mathbb{E}\left[f(w_{k+1}) - f(\zeta_k)\right] &\leq \left[-\eta + \frac{L\rho\eta^2}{2}\right]\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2} \end{aligned}$$

If $\eta \leq \frac{1}{\rho L}$,

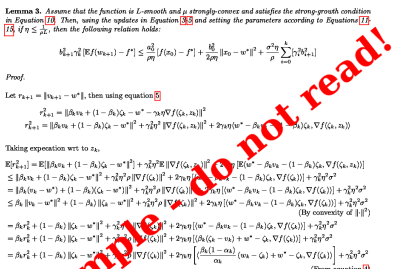
$$\begin{aligned} \mathbb{E}\left[f(w_{k+1}) - f(\zeta_k)\right] &\leq \left(\frac{-\eta}{2}\right)\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2} \\ \implies \left\|\nabla f(\zeta_k)\right\|^2 &\leq \left(\frac{2}{\eta}\right)\mathbb{E}[f(\zeta_k) - f(w_{k+1})] + L\eta\sigma^2 \end{aligned} \tag{17}$$

From equations 16 and 17

$$\begin{aligned} \mathbb{E}[r_{k+1}^2] &\leq \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + 2\gamma_k^2\rho\eta\mathbb{E}[f(\zeta_k) - f(w_{k+1})] \\ &\quad + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}(f(w_k) - f(\zeta_k)) + f^* - f(\zeta_k) - \frac{\mu}{2}\left\|\zeta_k - w^*\right\|^2\right] + \gamma_k^2\eta^2\sigma^2 + L\gamma_k^2\eta^2\sigma^2 \\ &= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + 2\gamma_k^2\eta\rho\mathbb{E}[f(\zeta_k) - f(w_{k+1})] \\ &\quad + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}(f(w_k) - f(\zeta_k)) + f^* - f(\zeta_k) - \frac{\mu}{2}\left\|\zeta_k - w^*\right\|^2\right] + 2\gamma_k^2\eta^2\sigma^2 \qquad \text{(Since } \eta \leq \tfrac{1}{\rho L}\text{)} \\ &= \beta_k r_k^2 + \left\|\zeta_k - w^*\right\|^2(1-\beta_k) - \gamma_k\mu\eta\right] + f(\zeta_k)\left[2\gamma_k^2\rho - 2\gamma_k\eta\cdot\frac{\beta_k(1-\alpha_k)}{\alpha_k} - 2\gamma_k\eta\right] \\ &\quad - 2\gamma_k^2\eta\mathbb{E}f(w_{k+1}) + 2\gamma_k\eta f^* + \left[2\gamma_k\eta\cdot\frac{\beta_k(1-\alpha_k)}{\alpha_k}\right]f(w_k) + 2\gamma_k^2\eta^2\sigma^2 \end{aligned}$$

Example - do not read!

Legitimate questions about performance analyses?



Lemma 3. *Assume that the function is L-smooth and μ strongly-convex and satisfies the strong-growth condition in Equation 10. Then, using the updates in Equation 3-5 and setting the parameters according to Equations 15-16 if $\eta \leq \frac{1}{\rho L}$, then the following relation holds:*

$$b_{k+1}^2 \eta \mathbb{E}\left[f(w_{k+1}) - f^*\right] \leq \frac{a_k^2}{\rho\eta}\left[f(x_0) - f^*\right] + \frac{b_0^2}{2\rho\eta}\left\|v_0 - w^*\right\|^2 + \frac{\alpha^2\eta}{\rho}\sum_{i=0}^{k}\left[\gamma_i^2 b_{i+1}^2\right]$$

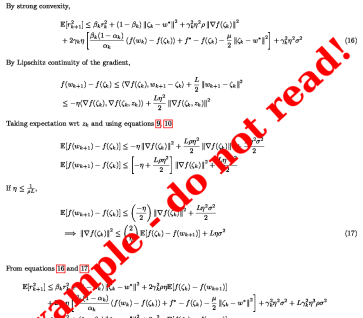
Proof.

Let $r_{k+1} = \|w_{k+1} - w^*\|$, then using equation 5

$$\begin{aligned} r_{k+1}^2 &= \|\beta_k v_k + (1-\beta_k)\zeta_k - w^* - \gamma_k \eta \nabla f(\zeta_k, z_k)\|^2 \\ r_{k+1}^2 &= \|\beta_k v_k + (1-\beta_k)\zeta_k - w^*\|^2 + \gamma_k^2\eta^2\left\|\nabla f(\zeta_k, z_k)\right\|^2 + 2\gamma_k\eta\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k, z_k)\rangle \end{aligned}$$

Taking expectation wrt to z_k,

$$\begin{aligned} \mathbb{E}[r_{k+1}^2] &= \mathbb{E}\left[\|\beta_k v_k + (1-\beta_k)\zeta_k - w^*\|^2\right] + \gamma_k^2\eta^2\mathbb{E}\left\|\nabla f(\zeta_k, z_k)\right\|^2 + 2\gamma_k\eta\mathbb{E}[\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k, z_k)\rangle] \\ &\leq \|\beta_k v_k + (1-\beta_k)\zeta_k - w^*\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle + \gamma_k^2\eta^2\sigma^2 \\ &= \|\beta_k(v_k - w^*) + (1-\beta_k)(\zeta_k - w^*)\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle + \gamma_k^2\eta^2\sigma^2 \\ &\leq \beta_k\|v_k - w^*\|^2 + (1-\beta_k)\|\zeta_k - w^*\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle + \gamma_k^2\eta^2\sigma^2 \\ &\qquad \text{(By convexity of } \|\cdot\|^2\text{)} \\ &= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\langle w^* - \beta_k v_k - (1-\beta_k)\zeta_k, \nabla f(\zeta_k)\rangle + \gamma_k^2\eta^2\sigma^2 \\ &= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}\langle w_k - \zeta_k\rangle + \langle w^* - \zeta_k, \nabla f(\zeta_k)\rangle\right] + \gamma_k^2\eta^2\sigma^2 \\ &\qquad \text{(From equation 4)} \\ &= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}\langle\nabla f(\zeta_k), (w_k - \zeta_k)\rangle + \langle\nabla f(\zeta_k), w^* - \zeta_k\rangle\right] + \gamma_k^2\eta^2\sigma^2 \\ &\leq \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}(f(w_k) - f(\zeta_k)) + \langle\nabla f(\zeta_k), w^* - \zeta_k\rangle\right] + \gamma_k^2\eta^2\sigma^2 \\ &\qquad \text{(By convexity)} \end{aligned}$$



By strong convexity,

$$\begin{aligned} \mathbb{E}[r_{k+1}^2] &\leq \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + \gamma_k^2\eta^2\rho\left\|\nabla f(\zeta_k)\right\|^2 \\ &\quad + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}(f(w_k) - f(\zeta_k)) + f^* - f(\zeta_k) - \frac{\mu}{2}\left\|\zeta_k - w^*\right\|^2\right] + \gamma_k^2\eta^2\sigma^2 \end{aligned} \tag{16}$$

By Lipschitz continuity of the gradient,

$$\begin{aligned} f(w_{k+1}) - f(\zeta_k) &\leq \langle\nabla f(\zeta_k), w_{k+1} - \zeta_k\rangle + \frac{L}{2}\left\|w_{k+1} - \zeta_k\right\|^2 \\ &\leq -\eta\langle\nabla f(\zeta_k), \nabla f(\zeta_k, z_k)\rangle + \frac{L\eta^2}{2}\left\|\nabla f(\zeta_k, z_k)\right\|^2 \end{aligned}$$

Taking expectation wrt z_k and using equations 9, 10

$$\begin{aligned} \mathbb{E}\left[f(w_{k+1}) - f(\zeta_k)\right] &\leq -\eta\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\rho\eta^2}{2}\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2} \\ \mathbb{E}\left[f(w_{k+1}) - f(\zeta_k)\right] &\leq \left[-\eta + \frac{L\rho\eta^2}{2}\right]\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2} \end{aligned}$$

If $\eta \leq \frac{1}{\rho L}$,

$$\begin{aligned} \mathbb{E}\left[f(w_{k+1}) - f(\zeta_k)\right] &\leq \left(\frac{-\eta}{2}\right)\left\|\nabla f(\zeta_k)\right\|^2 + \frac{L\eta^2\sigma^2}{2} \\ \implies \left\|\nabla f(\zeta_k)\right\|^2 &\leq \left(\frac{2}{\eta}\right)\mathbb{E}\left[f(\zeta_k) - f(w_{k+1})\right] + L\eta\sigma^2 \end{aligned} \tag{17}$$

From equations 16 and 17

$$\begin{aligned} \mathbb{E}[r_{k+1}^2] &\leq \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + 2\gamma_k^2\eta\rho\mathbb{E}[f(\zeta_k) - f(w_{k+1})] \\ &\quad + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}(f(w_k) - f(\zeta_k)) + f^* - f(\zeta_k) - \frac{\mu}{2}\left\|\zeta_k - w^*\right\|^2\right] + \gamma_k^2\eta^2\sigma^2 + L\gamma_k^2\eta^2\rho\sigma^2 \\ &= \beta_k r_k^2 + (1-\beta_k)\left\|\zeta_k - w^*\right\|^2 + 2\gamma_k^2\eta\rho\mathbb{E}[f(\zeta_k) - f(w_{k+1})] \\ &\quad + 2\gamma_k\eta\left[\frac{\beta_k(1-\alpha_k)}{\alpha_k}(f(w_k) - f(\zeta_k)) + f^* - f(\zeta_k) - \frac{\mu}{2}\left\|\zeta_k - w^*\right\|^2\right] + 2\gamma_k^2\eta^2\sigma^2 \qquad \text{(Since } \eta \leq \tfrac{1}{\rho L}\text{)} \\ &= \beta_k r_k^2 + \left\|\zeta_k - w^*\right\|^2\left[(1-\beta_k) - \gamma_k\mu\eta\right] + f(\zeta_k)\left[2\gamma_k^2\eta\rho - 2\gamma_k\eta\cdot\frac{\beta_k(1-\alpha_k)}{\alpha_k} - 2\gamma_k\eta\right] \\ &\quad - 2\gamma_k^2\eta\rho\mathbb{E}f(w_{k+1}) + 2\gamma_k\eta f^* + \left[2\gamma_k\eta\cdot\frac{\beta_k(1-\alpha_k)}{\alpha_k}\right]f(w_k) + 2\gamma_k^2\eta^2\sigma^2 \end{aligned}$$

- ⚠ Error-prone
- ⚠ Technical, lack global insights.
- ⚠ Few proof patterns.

- ? Easily fixable?
- ? Simple to adapt to variations of target inequality?
- ? Simple to adapt to algorithmic variations?

Convergence rate of a gradient step

Convergence rate of a gradient step

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

Convergence rate of a gradient step

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

Computing τ?[1]

[1] Original idea from [Drori and Teboulle, 2014]. Developments here from [T, Hendrickx, Glineur, 2017].

Convergence rate of a gradient step

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

Computing τ?[1]

$$\tau = \max_{f, \mathbf{x_0}, \mathbf{x_1}, x_\star} \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2}$$

$$\text{s.t. } f \in \mathcal{F}_{\mu,L}$$

Functional class

[1] Original idea from [Drori and Teboulle, 2014]. Developments here from [T, Hendrickx, Glineur, 2017].

Convergence rate of a gradient step

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

Computing τ?[1]

$$\tau = \max_{f, x_0, x_1, x_\star} \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2}$$

$$\text{s.t. } f \in \mathcal{F}_{\mu,L} \qquad \text{Functional class}$$

$$x_1 = x_0 - \alpha \nabla f(x_0) \qquad \text{Algorithm}$$

$$\nabla f(x_\star) = 0 \qquad \text{Optimality of } x_\star$$

[1] Original idea from [Drori and Teboulle, 2014]. Developments here from [T, Hendrickx, Glineur, 2017].

Convergence rate of a gradient step

Toy example: What is the smallest τ such that:

$$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2,$$

for all

- $d \in \mathbb{N}$, L-smooth and μ-strongly convex function f (notation $f \in \mathcal{F}_{\mu,L}$),
- x_0, and x_1 generated by gradient step $x_1 = x_0 - \alpha \nabla f(x_0)$,
- $x_\star = \operatorname*{argmin}_x f(x)$?

Computing τ?[1]

$$\tau = \max_{f, x_0, x_1, x_\star} \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2}$$

$$\text{s.t. } f \in \mathcal{F}_{\mu,L} \qquad \text{Functional class}$$

$$x_1 = x_0 - \alpha \nabla f(x_0) \qquad \text{Algorithm}$$

$$\nabla f(x_\star) = 0 \qquad \text{Optimality of } x_\star$$

Variables: f, x_0, x_1, $x_\star$.

Parameters: μ, L, α.

[1] Original idea from [Drori and Teboulle, 2014]. Developments here from [T, Hendrickx, Glineur, 2017].

Sampled version

Sampled version

◇ Performance estimation problem :

$$\begin{aligned} \max_{f, x_0, x_1, x_\star} \quad & \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2} \\ \text{subject to} \quad & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex}, \\ & x_1 = x_0 - \alpha \nabla f(x_0) \\ & \nabla f(x_\star) = 0. \end{aligned}$$

Sampled version

◇ Performance estimation problem **(Variables:** f**,** x_0**,** x_1**,** $x_\star$**)**:

$$\begin{aligned} \max_{f, x_0, x_1, x_\star} \quad & \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2} \\ \text{subject to} \quad & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex}, \\ & x_1 = x_0 - \alpha \nabla f(x_0) \\ & \nabla f(x_\star) = 0. \end{aligned}$$

Sampled version

◇ Performance estimation problem **(Variables:** f**,** x_0**,** x_1**,** $x_\star$**)**:

$$\begin{aligned} \max_{f, x_0, x_1, x_\star} \quad & \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2} \\ \text{subject to} \quad & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex}, \\ & x_1 = x_0 - \alpha \nabla f(x_0) \\ & \nabla f(x_\star) = 0. \end{aligned}$$

◇ Sampled version:

Sampled version

◇ Performance estimation problem **(Variables:** f, x_0, x_1, $x_\star$**)**:

$$\begin{aligned} \max_{f,x_0,x_1,x_\star} \quad & \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2} \\ \text{subject to} \quad & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex}, \\ & x_1 = x_0 - \alpha \nabla f(x_0) \\ & \nabla f(x_\star) = 0. \end{aligned}$$

◇ Sampled version:

$$\begin{aligned} \max_{\substack{x_0,x_1,x_\star \\ g_0,g_\star \\ f_0,f_\star}} \quad & \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2} \\ \text{subject to} \quad & \exists f \in \mathcal{F}_{\mu,L} \text{ such that } \begin{cases} f_i = f(x_i) & i = 0,\star \\ g_i = \nabla f(x_i) & i = 0,\star \end{cases} \\ & x_1 = x_0 - \alpha g_0 \\ & g_\star = 0. \end{aligned}$$

Sampled version

◇ Performance estimation problem **(Variables:** f, x_0, x_1, $x_\star$**)**:

$$\begin{aligned} \max_{f,x_0,x_1,x_\star} \quad & \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2} \\ \text{subject to} \quad & f \text{ is } L\text{-smooth and } \mu\text{-strongly convex}, \\ & x_1 = x_0 - \alpha \nabla f(x_0) \\ & \nabla f(x_\star) = 0. \end{aligned}$$

◇ Sampled version: **(Variables:** x_0, x_1, $x_\star$, g_0, $g_\star$, f_0, $f_\star$**)**

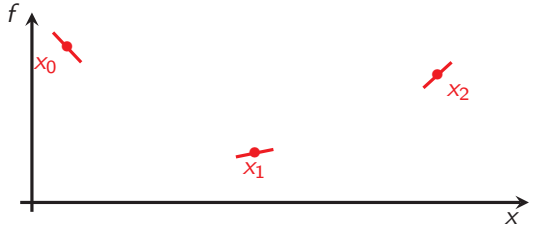
$$\begin{aligned} \max_{\substack{x_0,x_1,x_\star \\ g_0,g_\star \\ f_0,f_\star}} \quad & \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2} \\ \text{subject to} \quad & \exists f \in \mathcal{F}_{\mu,L} \text{ such that } \begin{cases} f_i = f(x_i) & i = 0, \star \\ g_i = \nabla f(x_i) & i = 0, \star \end{cases} \\ & x_1 = x_0 - \alpha g_0 \\ & g_\star = 0. \end{aligned}$$

Smooth strongly convex interpolation (or extension)

Let I index set, and associated $\{(x_i, g_i, f_i)\}_{i \in I}$: points x_i, (sub)gradients g_i and function values f_i.

Smooth strongly convex interpolation (or extension)

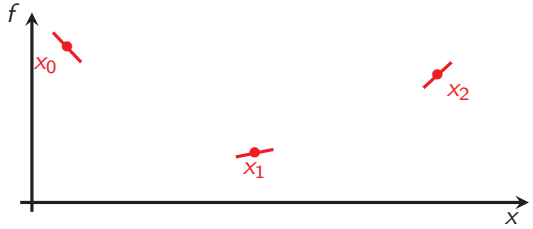
Let I index set, and associated $\{(x_i, g_i, f_i)\}_{i \in I}$: points x_i, (sub)gradients g_i and function values f_i.



? Possible to find $f \in \mathcal{F}_{\mu,L}$ such that $f(x_i) = f_i$, and $g_i = \nabla f(x_i)$ $\forall i \in I$?

Smooth strongly convex interpolation (or extension)

Let I index set, and associated $\{(x_i, g_i, f_i)\}_{i\in I}$: points x_i, (sub)gradients g_i and function values f_i.



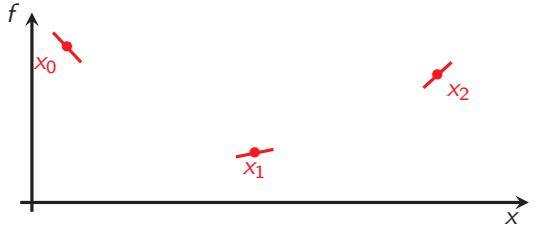
? Possible to find $f \in \mathcal{F}_{\mu,L}$ such that $f(x_i) = f_i$, and $g_i = \nabla f(x_i)\ \forall i \in I$?

- Necessary and sufficient condition: $\forall i, j \in I$

$$f_i \geqslant f_j + \langle g_j, x_i - x_j\rangle + \tfrac{1}{2L}\|g_i - g_j\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_i - x_j - \tfrac{1}{L}(g_i - g_j)\right\|^2.$$

Smooth strongly convex interpolation (or extension)

Let I index set, and associated $\{(x_i, g_i, f_i)\}_{i\in I}$: points x_i, (sub)gradients g_i and function values f_i.



? Possible to find $f \in \mathcal{F}_{\mu,L}$ such that $f(x_i) = f_i$, and $g_i = \nabla f(x_i)\ \forall i \in I$?

- Necessary and sufficient condition: $\forall i, j \in I$

$$f_i \geqslant f_j + \langle g_j, x_i - x_j\rangle + \tfrac{1}{2L}\|g_i - g_j\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_i - x_j - \tfrac{1}{L}(g_i - g_j)\right\|^2.$$

- Simpler example: pick $\mu = 0$ and $L = \infty$ (just convexity):

$$f_i \geqslant f_j + \langle g_j, x_i - x_j\rangle.$$

Replace constraints

Replace constraints

◇ Interpolation conditions allow removing red constraints

$$\max_{\substack{x_0, x_1, x_\star \\ g_0, g_\star \\ f_0, f_\star}} \quad \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2}$$

subject to $\quad \exists f \in \mathcal{F}_{\mu,L}$ such that $\begin{cases} f_i = f(x_i) & i = 0, \star \\ g_i = \nabla f(x_i) & i = 0, \star \end{cases}$

$$x_1 = x_0 - \alpha g_0$$
$$g_\star = 0,$$

Replace constraints

◇ Interpolation conditions allow removing red constraints

$$\max_{\substack{x_0,x_1,x_\star \\ g_0,g_\star \\ f_0,f_\star}} \quad \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2}$$

$$\text{subject to} \quad \exists f \in \mathcal{F}_{\mu,L} \text{ such that } \begin{cases} f_i = f(x_i) & i = 0,\star \\ g_i = \nabla f(x_i) & i = 0,\star \end{cases}$$

$$x_1 = x_0 - \alpha g_0$$

$$g_\star = 0,$$

◇ replacing them by

$$f_\star \geqslant f_0 + \langle g_0, x_\star - x_0\rangle + \tfrac{1}{2L}\|g_\star - g_0\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_\star - x_0 - \tfrac{1}{L}(g_\star - g_0)\right\|^2$$

$$f_0 \geqslant f_\star + \langle g_\star, x_0 - x_\star\rangle + \tfrac{1}{2L}\|g_0 - g_\star\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}(g_0 - g_\star)\right\|^2.$$

Replace constraints

◇ Interpolation conditions allow removing red constraints

$$\max_{\substack{x_0,x_1,x_\star \\ g_0,g_\star \\ f_0,f_\star}} \quad \frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2}$$

$$\text{subject to} \quad \exists f \in \mathcal{F}_{\mu,L} \text{ such that } \begin{cases} f_i = f(x_i) & i = 0,\star \\ g_i = \nabla f(x_i) & i = 0,\star \end{cases}$$

$$x_1 = x_0 - \alpha g_0$$

$$g_\star = 0,$$

◇ replacing them by

$$f_\star \geqslant f_0 + \langle g_0, x_\star - x_0\rangle + \tfrac{1}{2L}\|g_\star - g_0\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_\star - x_0 - \tfrac{1}{L}(g_\star - g_0)\right\|^2$$

$$f_0 \geqslant f_\star + \langle g_\star, x_0 - x_\star\rangle + \tfrac{1}{2L}\|g_0 - g_\star\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}(g_0 - g_\star)\right\|^2.$$

◇ Same optimal value (no relaxation): non-convex quadratic problem.

Semidefinite lifting

- Define $P \triangleq [x_0 - x_\star,\, g_0] \in \mathbb{R}^{d\times 2}$ and $F \triangleq f_0 - f_\star$.

Semidefinite lifting

- Define $P \triangleq [x_0 - x_\star,\ g_0] \in \mathbb{R}^{d\times 2}$ and $F \triangleq f_0 - f_\star$.
- Using new variables $G \succcurlyeq 0$ and F

$$G \triangleq P^T P = \begin{bmatrix} \|x_0 - x_\star\|^2 & \langle g_0, x_0 - x_\star\rangle \\ \langle g_0, x_0 - x_\star\rangle & \|g_0\|^2 \end{bmatrix} \succcurlyeq 0,$$

Semidefinite lifting

- Define $P \triangleq [x_0 - x_\star,\ g_0] \in \mathbb{R}^{d\times 2}$ and $F \triangleq f_0 - f_\star$.
- Using new variables $G \succcurlyeq 0$ and F

$$G \triangleq P^T P = \begin{bmatrix} \|x_0 - x_\star\|^2 & \langle g_0, x_0 - x_\star\rangle \\ \langle g_0, x_0 - x_\star\rangle & \|g_0\|^2 \end{bmatrix} \succcurlyeq 0,$$

- previous problem can be relaxed to 2×2 SDP

$$\begin{aligned} \max_{G,\,F} \quad & G_{1,1} + \alpha^2 G_{2,2} - 2\alpha G_{1,2} \\ \text{subject to} \quad & F + \tfrac{L\mu}{2(L-\mu)} G_{1,1} + \tfrac{1}{2(L-\mu)} G_{2,2} - \tfrac{L}{L-\mu} G_{1,2} \leqslant 0 \\ & -F + \tfrac{L\mu}{2(L-\mu)} G_{1,1} + \tfrac{1}{2(L-\mu)} G_{2,2} - \tfrac{\mu}{L-\mu} G_{1,2} \leqslant 0 \\ & G_{1,1} = 1 \\ & G \succcurlyeq 0 \end{aligned}$$

Semidefinite lifting

- Define $P \triangleq [x_0 - x_\star,\, g_0] \in \mathbb{R}^{d\times 2}$ and $F \triangleq f_0 - f_\star$.
- Using new variables $G \succcurlyeq 0$ and F

$$G \triangleq P^T P = \begin{bmatrix} \|x_0 - x_\star\|^2 & \langle g_0, x_0 - x_\star\rangle \\ \langle g_0, x_0 - x_\star\rangle & \|g_0\|^2 \end{bmatrix} \succcurlyeq 0,$$

- previous problem can be relaxed to 2×2 SDP

$$\begin{aligned} \max_{G,\,F} \quad & G_{1,1} + \alpha^2 G_{2,2} - 2\alpha G_{1,2} \\ \text{subject to} \quad & F + \tfrac{L\mu}{2(L-\mu)} G_{1,1} + \tfrac{1}{2(L-\mu)} G_{2,2} - \tfrac{L}{L-\mu} G_{1,2} \leqslant 0 \\ & -F + \tfrac{L\mu}{2(L-\mu)} G_{1,1} + \tfrac{1}{2(L-\mu)} G_{2,2} - \tfrac{\mu}{L-\mu} G_{1,2} \leqslant 0 \\ & G_{1,1} = 1 \\ & G \succcurlyeq 0 \end{aligned}$$

(using homogeneity argument and substituting x_1 and $g_\star$).

Semidefinite lifting

- Define $P \triangleq [x_0 - x_\star,\ g_0] \in \mathbb{R}^{d\times 2}$ and $F \triangleq f_0 - f_\star$.
- Using new variables $G \succcurlyeq 0$ and F

$$G \triangleq P^T P = \begin{bmatrix} \|x_0 - x_\star\|^2 & \langle g_0, x_0 - x_\star\rangle \\ \langle g_0, x_0 - x_\star\rangle & \|g_0\|^2 \end{bmatrix} \succcurlyeq 0,$$

- previous problem can be relaxed to 2×2 SDP

$$\begin{aligned} \max_{G,\,F} \quad & G_{1,1} + \alpha^2 G_{2,2} - 2\alpha G_{1,2} \\ \text{subject to} \quad & F + \tfrac{L\mu}{2(L-\mu)} G_{1,1} + \tfrac{1}{2(L-\mu)} G_{2,2} - \tfrac{L}{L-\mu} G_{1,2} \leqslant 0 \\ & -F + \tfrac{L\mu}{2(L-\mu)} G_{1,1} + \tfrac{1}{2(L-\mu)} G_{2,2} - \tfrac{\mu}{L-\mu} G_{1,2} \leqslant 0 \\ & G_{1,1} = 1 \\ & G \succcurlyeq 0 \end{aligned}$$

  (using homogeneity argument and substituting x_1 and $g_\star$).
- Assuming $x_0, x_\star, g_0 \in \mathbb{R}^d$ with $d \geqslant 2$, same optimal value as original problem!

Semidefinite lifting

- Define $P \triangleq [x_0 - x_\star,\ g_0] \in \mathbb{R}^{d\times 2}$ and $F \triangleq f_0 - f_\star$.
- Using new variables $G \succcurlyeq 0$ and F

$$G \triangleq P^T P = \begin{bmatrix} \|x_0 - x_\star\|^2 & \langle g_0, x_0 - x_\star\rangle \\ \langle g_0, x_0 - x_\star\rangle & \|g_0\|^2 \end{bmatrix} \succcurlyeq 0,$$

- previous problem can be relaxed to 2×2 SDP

$$\begin{aligned} \max_{G,\,F} \quad & G_{1,1} + \alpha^2 G_{2,2} - 2\alpha G_{1,2} \\ \text{subject to} \quad & F + \tfrac{L\mu}{2(L-\mu)} G_{1,1} + \tfrac{1}{2(L-\mu)} G_{2,2} - \tfrac{L}{L-\mu} G_{1,2} \leqslant 0 \\ & -F + \tfrac{L\mu}{2(L-\mu)} G_{1,1} + \tfrac{1}{2(L-\mu)} G_{2,2} - \tfrac{\mu}{L-\mu} G_{1,2} \leqslant 0 \\ & G_{1,1} = 1 \\ & G \succcurlyeq 0 \end{aligned}$$

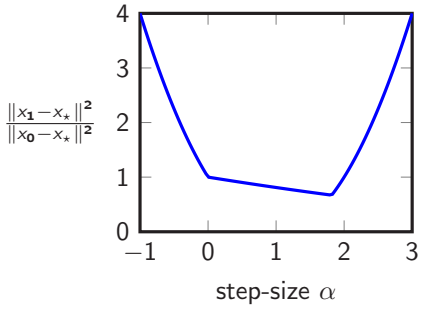
  (using homogeneity argument and substituting x_1 and $g_\star$).
- Assuming $x_0, x_\star, g_0 \in \mathbb{R}^d$ with $d \geqslant 2$, same optimal value as original problem!
- For $d = 1$ same as original problem by adding $\operatorname{rank}(G) \leqslant 1$.

Numerical solution of the SDP

Fix $L = 1$, $\mu = .1$ and solve the SDP for a few values of α.

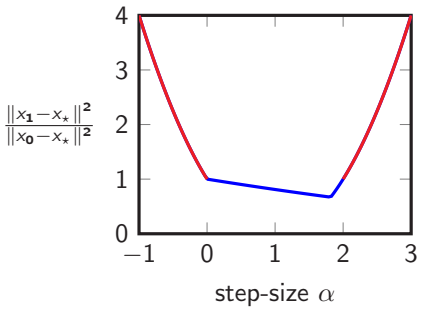
Numerical solution of the SDP

Fix $L = 1$, $\mu = .1$ and solve the SDP for a few values of α.



Numerical solution of the SDP

Fix $L = 1$, $\mu = .1$ and solve the SDP for a few values of α.



Observations:

- ◇ numerics match the known $\max\{(1-\alpha L)^2, (1-\alpha\mu)^2\}$
- ◇ recovers that gradient descent converges for $\alpha \in (0, 2/L)$ (divergence otherwise).

Dual problem

- Dual problem is

$$\min_{\tau,\lambda_1,\lambda_2 \geqslant 0} \tau$$

$$\text{subject to } S = \begin{bmatrix} \tau - 1 + \frac{\lambda_1 L\mu}{L-\mu} & \alpha - \frac{\lambda_1(L+\mu)}{2(L-\mu)} \\ \alpha - \frac{\lambda_1(L+\mu)}{2(L-\mu)} & \frac{\lambda_1}{L-\mu} - \alpha^2 \end{bmatrix} \succcurlyeq 0$$

$$0 = \lambda_1 - \lambda_2.$$

Dual problem

◇ Dual problem is

$$\min_{\tau,\lambda_1,\lambda_2 \geqslant 0} \tau$$
$$\text{subject to } S = \begin{bmatrix} \tau - 1 + \frac{\lambda_1 L\mu}{L-\mu} & \alpha - \frac{\lambda_1(L+\mu)}{2(L-\mu)} \\ \alpha - \frac{\lambda_1(L+\mu)}{2(L-\mu)} & \frac{\lambda_1}{L-\mu} - \alpha^2 \end{bmatrix} \succcurlyeq 0$$
$$0 = \lambda_1 - \lambda_2.$$

◇ Weak duality: any dual feasible point $\equiv$ valid worst-case convergence rate

Dual problem

- Dual problem is

$$\min_{\tau,\lambda_1,\lambda_2 \geqslant 0} \tau$$
$$\text{subject to } S = \begin{bmatrix} \tau - 1 + \frac{\lambda_1 L\mu}{L-\mu} & \alpha - \frac{\lambda_1(L+\mu)}{2(L-\mu)} \\ \alpha - \frac{\lambda_1(L+\mu)}{2(L-\mu)} & \frac{\lambda_1}{L-\mu} - \alpha^2 \end{bmatrix} \succcurlyeq 0$$
$$0 = \lambda_1 - \lambda_2.$$

- Weak duality: any dual feasible point $\equiv$ valid worst-case convergence rate
- Direct consequence: for any $\tau \geqslant 0$ we have

$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2$ for all $f \in \mathcal{F}_{\mu,L}$, all $x_0 \in \mathbb{R}^d$, all $d \in \mathbb{N}$, with $x_1 = x_0 - \alpha \nabla f(x_0)$.

$$\Uparrow$$

$$\exists \lambda \geqslant 0 : \begin{bmatrix} \tau - 1 + \frac{\lambda L\mu}{L-\mu} & \alpha - \frac{\lambda(L+\mu)}{2(L-\mu)} \\ \alpha - \frac{\lambda(L+\mu)}{2(L-\mu)} & \frac{\lambda}{L-\mu} - \alpha^2 \end{bmatrix} \succcurlyeq 0$$

Dual problem

- Dual problem is

$$\min_{\tau,\lambda_1,\lambda_2 \geqslant 0} \tau$$
$$\text{subject to } S = \begin{bmatrix} \tau - 1 + \frac{\lambda_1 L\mu}{L-\mu} & \alpha - \frac{\lambda_1(L+\mu)}{2(L-\mu)} \\ \alpha - \frac{\lambda_1(L+\mu)}{2(L-\mu)} & \frac{\lambda_1}{L-\mu} - \alpha^2 \end{bmatrix} \succcurlyeq 0$$
$$0 = \lambda_1 - \lambda_2.$$

- Weak duality: any dual feasible point $\equiv$ valid worst-case convergence rate ($\Uparrow$).
- Direct consequence: for any $\tau \geqslant 0$ we have

$\|x_1 - x_\star\|^2 \leqslant \tau \|x_0 - x_\star\|^2$ for all $f \in \mathcal{F}_{\mu,L}$, all $x_0 \in \mathbb{R}^d$, all $d \in \mathbb{N}$, with $x_1 = x_0 - \alpha \nabla f(x_0)$.

$\Uparrow$

$$\exists \lambda \geqslant 0 : \begin{bmatrix} \tau - 1 + \frac{\lambda L\mu}{L-\mu} & \alpha - \frac{\lambda(L+\mu)}{2(L-\mu)} \\ \alpha - \frac{\lambda(L+\mu)}{2(L-\mu)} & \frac{\lambda}{L-\mu} - \alpha^2 \end{bmatrix} \succcurlyeq 0$$

Dual problem

- Dual problem is

$$\begin{aligned}&\min_{\tau,\lambda_1,\lambda_2\geqslant 0} \tau\\ &\text{subject to } S=\begin{bmatrix}\tau-1+\frac{\lambda_1 L\mu}{L-\mu} & \alpha-\frac{\lambda_1(L+\mu)}{2(L-\mu)}\\ \alpha-\frac{\lambda_1(L+\mu)}{2(L-\mu)} & \frac{\lambda_1}{L-\mu}-\alpha^2\end{bmatrix}\succcurlyeq 0\\ &\qquad\qquad 0=\lambda_1-\lambda_2.\end{aligned}$$

- Weak duality: any dual feasible point $\equiv$ valid worst-case convergence rate ($\Uparrow$).
- Direct consequence: for any $\tau\geqslant 0$ we have

$$\|x_1-x_\star\|^2\leqslant\tau\|x_0-x_\star\|^2 \text{ for all } f\in\mathcal{F}_{\mu,L}, \text{ all } x_0\in\mathbb{R}^d, \text{ all } d\in\mathbb{N}, \text{ with } x_1=x_0-\alpha\nabla f(x_0).$$

$$\Uparrow$$

$$\exists\lambda\geqslant 0:\begin{bmatrix}\tau-1+\frac{\lambda L\mu}{L-\mu} & \alpha-\frac{\lambda(L+\mu)}{2(L-\mu)}\\ \alpha-\frac{\lambda(L+\mu)}{2(L-\mu)} & \frac{\lambda}{L-\mu}-\alpha^2\end{bmatrix}\succcurlyeq 0$$

- Strong duality holds ($\exists$ Slater point): any valid worst-case convergence rate $\equiv$ valid dual feasible point ($\Downarrow$).

Dual problem

- Dual problem is

$$\begin{aligned}&\min_{\tau,\lambda_1,\lambda_2\geqslant 0} \tau\\ &\text{subject to } S = \begin{bmatrix} \tau - 1 + \frac{\lambda_1 L\mu}{L-\mu} & \alpha - \frac{\lambda_1(L+\mu)}{2(L-\mu)} \\ \alpha - \frac{\lambda_1(L+\mu)}{2(L-\mu)} & \frac{\lambda_1}{L-\mu} - \alpha^2 \end{bmatrix} \succcurlyeq 0\\ &\qquad\quad 0 = \lambda_1 - \lambda_2.\end{aligned}$$

- Weak duality: any dual feasible point $\equiv$ valid worst-case convergence rate ($\Uparrow$).
- Direct consequence: for any $\tau \geqslant 0$ we have

$$\|x_1 - x_\star\|^2 \leqslant \tau\|x_0 - x_\star\|^2 \text{ for all } f \in \mathcal{F}_{\mu,L}, \text{ all } x_0 \in \mathbb{R}^d, \text{ all } d \in \mathbb{N}, \text{ with } x_1 = x_0 - \alpha\nabla f(x_0).$$

$$\Uparrow \quad \Downarrow$$

$$\exists \lambda \geqslant 0 : \begin{bmatrix} \tau - 1 + \frac{\lambda L\mu}{L-\mu} & \alpha - \frac{\lambda(L+\mu)}{2(L-\mu)} \\ \alpha - \frac{\lambda(L+\mu)}{2(L-\mu)} & \frac{\lambda}{L-\mu} - \alpha^2 \end{bmatrix} \succcurlyeq 0$$

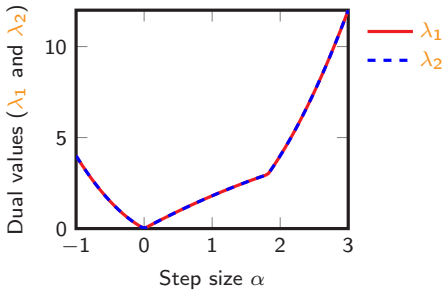
- Strong duality holds ($\exists$ Slater point): any valid worst-case convergence rate $\equiv$ valid dual feasible point ($\Downarrow$).

Dual solutions

Fix $L = 1$, $\mu = .1$ and solve the dual SDP for a few values of α.

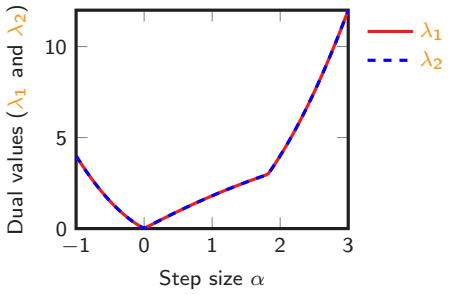
Dual solutions

Fix $L = 1$, $\mu = .1$ and solve the dual SDP for a few values of α.



Dual solutions

Fix $L = 1$, $\mu = .1$ and solve the dual SDP for a few values of α.



Numerics match $\lambda_1 = \lambda_2 = 2|\alpha|\rho(\alpha)$ with $\rho(\alpha) = \max\{\alpha L - 1, 1 - \alpha\mu\}$.

Recovering a "standard" proof

Gradient with $\alpha = \frac{1}{L}$. Perform weighted sum of two inequalities

Recovering a "standard" proof

Gradient with $\alpha = \frac{1}{L}$. Perform weighted sum of two inequalities

$$f_0 \geqslant f_\star + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\|^2 \qquad : \lambda_1$$

$$f_\star \geqslant f_0 + \langle \nabla f(x_0), x_\star - x_0\rangle + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\|^2 \qquad : \lambda_2$$

Recovering a "standard" proof

Gradient with $\alpha = \frac{1}{L}$. Perform weighted sum of two inequalities

$$f_0 \geqslant f_\star + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\|^2 \qquad : \lambda_1$$

$$f_\star \geqslant f_0 + \langle \nabla f(x_0), x_\star - x_0\rangle + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\|^2 \qquad : \lambda_2$$

with $\lambda_1, \lambda_2 \geqslant 0$. Weighted sum can be reformulated as

Recovering a "standard" proof

Gradient with $\alpha = \frac{1}{L}$. Perform weighted sum of two inequalities

$$f_0 \geqslant f_\star + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\|^2 \qquad : \lambda_1 = 2\alpha(1-\alpha\mu)$$

$$f_\star \geqslant f_0 + \langle \nabla f(x_0), x_\star - x_0\rangle + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\|^2 \qquad : \lambda_2 = 2\alpha(1-\alpha\mu)$$

with $\lambda_1, \lambda_2 \geqslant 0$. Weighted sum can be reformulated as

Recovering a "standard" proof

Gradient with $\alpha = \frac{1}{L}$. Perform weighted sum of two inequalities

$$f_0 \geqslant f_\star + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\right\|^2 \qquad : \lambda_1 = 2\alpha(1-\alpha\mu)$$

$$f_\star \geqslant f_0 + \langle \nabla f(x_0), x_\star - x_0\rangle + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\right\|^2 \qquad : \lambda_2 = 2\alpha(1-\alpha\mu)$$

with $\lambda_1, \lambda_2 \geqslant 0$. Weighted sum can be reformulated as

$$\|x_1 - x_\star\|^2 \leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2 - \underbrace{\alpha\tfrac{2-\alpha(L+\mu)}{L-\mu}\|\mu(x_0 - x_\star) - \nabla f(x_0)\|^2}\,,$$

Recovering a "standard" proof

Gradient with $\alpha = \frac{1}{L}$. Perform weighted sum of two inequalities

$$f_0 \geqslant f_\star + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\right\|^2 \qquad : \lambda_1 = 2\alpha(1-\alpha\mu)$$

$$f_\star \geqslant f_0 + \langle \nabla f(x_0), x_\star - x_0\rangle + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\right\|^2 \qquad : \lambda_2 = 2\alpha(1-\alpha\mu)$$

with $\lambda_1, \lambda_2 \geqslant 0$. Weighted sum can be reformulated as

$$\|x_1 - x_\star\|^2 \leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2 - \underbrace{\alpha\tfrac{2-\alpha(L+\mu)}{L-\mu}\|\mu(x_0 - x_\star) - \nabla f(x_0)\|^2}_{\geqslant 0},$$

Recovering a "standard" proof

Gradient with $\alpha = \frac{1}{L}$. Perform weighted sum of two inequalities

$$f_0 \geqslant f_\star + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\right\|^2 \qquad : \lambda_1 = 2\alpha(1-\alpha\mu)$$

$$f_\star \geqslant f_0 + \langle \nabla f(x_0), x_\star - x_0\rangle + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\right\|^2 \qquad : \lambda_2 = 2\alpha(1-\alpha\mu)$$

with $\lambda_1, \lambda_2 \geqslant 0$. Weighted sum can be reformulated as

$$\begin{aligned}\|x_1 - x_\star\|^2 &\leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2 - \underbrace{\alpha\tfrac{2-\alpha(L+\mu)}{L-\mu}\|\mu(x_0 - x_\star) - \nabla f(x_0)\|^2}_{\geqslant 0},\\ &\leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2,\end{aligned}$$

Recovering a "standard" proof

Gradient with $\alpha = \frac{1}{L}$. Perform weighted sum of two inequalities

$$f_0 \geqslant f_\star + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\|^2 \qquad : \lambda_1 = 2\alpha(1-\alpha\mu)$$
$$f_\star \geqslant f_0 + \langle \nabla f(x_0), x_\star - x_0\rangle + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\|^2 \qquad : \lambda_2 = 2\alpha(1-\alpha\mu)$$

with $\lambda_1, \lambda_2 \geqslant 0$. Weighted sum can be reformulated as

$$\begin{aligned}\|x_1 - x_\star\|^2 &\leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2 - \underbrace{\alpha\tfrac{2-\alpha(L+\mu)}{L-\mu}\|\mu(x_0 - x_\star) - \nabla f(x_0)\|^2}_{\geqslant 0}, \\ &\leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2,\end{aligned}$$

leading to $\|x_1 - x_\star\|^2 \leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2$

Recovering a "standard" proof

Gradient with $\alpha = \frac{1}{L}$. Perform weighted sum of two inequalities

$$f_0 \geqslant f_\star + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\right\|^2 \qquad : \lambda_1 = 2\alpha(1-\alpha\mu)$$

$$f_\star \geqslant f_0 + \langle \nabla f(x_0), x_\star - x_0\rangle + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\right\|^2 \qquad : \lambda_2 = 2\alpha(1-\alpha\mu)$$

with $\lambda_1, \lambda_2 \geqslant 0$. Weighted sum can be reformulated as

$$\begin{aligned}\|x_1 - x_\star\|^2 &\leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2 \underbrace{- \alpha\tfrac{2-\alpha(L+\mu)}{L-\mu}\|\mu(x_0 - x_\star) - \nabla f(x_0)\|^2}_{\geqslant 0,\text{ or } = 0 \text{ when worst-case is achieved}},\\ &\leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2,\end{aligned}$$

leading to $\|x_1 - x_\star\|^2 \leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2$

Recovering a "standard" proof

Gradient with $\alpha = \frac{1}{L}$. Perform weighted sum of two inequalities

$$f_0 \geqslant f_\star + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\right\|^2 \qquad : \lambda_1 = 2\alpha(1-\alpha\mu)$$

$$f_\star \geqslant f_0 + \langle \nabla f(x_0), x_\star - x_0\rangle + \tfrac{1}{2L}\|\nabla f(x_0)\|^2 + \tfrac{\mu}{2(1-\mu/L)}\left\|x_0 - x_\star - \tfrac{1}{L}\nabla f(x_0)\right\|^2 \qquad : \lambda_2 = 2\alpha(1-\alpha\mu)$$

with $\lambda_1, \lambda_2 \geqslant 0$. Weighted sum can be reformulated as

$$\begin{aligned}\|x_1 - x_\star\|^2 &\leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2 \underbrace{- \alpha\tfrac{2-\alpha(L+\mu)}{L-\mu}\|\mu(x_0 - x_\star) - \nabla f(x_0)\|^2}_{\geqslant 0,\text{ or } = 0 \text{ when worst-case is achieved}},\\ &\leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2,\end{aligned}$$

leading to $\|x_1 - x_\star\|^2 \leqslant (1-\alpha\mu)^2\|x_0 - x_\star\|^2$ (tight).

What did we do, so far?

Summary:

What did we do, so far?

Summary:

- ◇ we computed the smallest $\tau(\mu, L, \alpha)$ such that

$$\|x_1 - x_\star\|^2 \leqslant \tau(\mu, L, \alpha) \|x_0 - x_\star\|^2$$

is satisfied for all $x_0 \in \mathbb{R}^d$, $d \in \mathbb{N}$, $f \in \mathcal{F}_{\mu,L}$, and $x_1 = x_0 - \alpha \nabla f(x_0)$.

What did we do, so far?

Summary:

- ◇ we computed the smallest $\tau(\mu, L, \alpha)$ such that

$$\|x_1 - x_\star\|^2 \leqslant \tau(\mu, L, \alpha) \|x_0 - x_\star\|^2$$

  is satisfied for all $x_0 \in \mathbb{R}^d$, $d \in \mathbb{N}$, $f \in \mathcal{F}_{\mu,L}$, and $x_1 = x_0 - \alpha \nabla f(x_0)$.

- ◇ Feasible points to primal SDP correspond to lower bounds on $\tau(\mu, L, \alpha)$.

What did we do, so far?

Summary:

- ◇ we computed the smallest $\tau(\mu, L, \alpha)$ such that

$$\|x_1 - x_\star\|^2 \leqslant \tau(\mu, L, \alpha) \|x_0 - x_\star\|^2$$

  is satisfied for all $x_0 \in \mathbb{R}^d$, $d \in \mathbb{N}$, $f \in \mathcal{F}_{\mu,L}$, and $x_1 = x_0 - \alpha \nabla f(x_0)$.
- ◇ Feasible points to primal SDP correspond to lower bounds on $\tau(\mu, L, \alpha)$.
- ◇ Feasible points to dual SDP correspond to upper bounds on $\tau(\mu, L, \alpha)$.

What did we do, so far?

Summary:

- ◇ we computed the smallest $\tau(\mu, L, \alpha)$ such that

$$\|x_1 - x_\star\|^2 \leqslant \tau(\mu, L, \alpha) \|x_0 - x_\star\|^2$$

  is satisfied for all $x_0 \in \mathbb{R}^d$, $d \in \mathbb{N}$, $f \in \mathcal{F}_{\mu,L}$, and $x_1 = x_0 - \alpha \nabla f(x_0)$.

- ◇ Feasible points to primal SDP correspond to lower bounds on $\tau(\mu, L, \alpha)$.
- ◇ Feasible points to dual SDP correspond to upper bounds on $\tau(\mu, L, \alpha)$.
  - proof via linear combinations of interpolation inequalities (evaluated at $\{x_k\}_k$ and $x_\star$),
  - proofs can be rewritten as a "sum-of-squares" certificates (sum of squared norms).

Why/when does it work?

The methodology applies, as is, as soon as:

Why/when does it work?

The methodology applies, as is, as soon as:

- ◇ performance measure and initial condition are linear in G and F,

Why/when does it work?

The methodology applies, as is, as soon as:

- performance measure and initial condition are linear in G and F,
- interpolation inequalities are linear in G and F,

Why/when does it work?

The methodology applies, as is, as soon as:

- ◇ performance measure and initial condition are linear in G and F,
- ◇ interpolation inequalities are linear in G and F,
- ◇ algorithm can be described linearly in G and F

Why/when does it work?

The methodology applies, as is, as soon as:

- ◇ performance measure and initial condition are linear in G and F,
- ◇ interpolation inequalities are linear in G and F,
- ◇ algorithm can be described linearly in G and F

(but other cases are not doomed).

Insight into proof structures?

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable) with $f \in \mathcal{F}$. Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x).$$

For example, assume ($x_\star = 0$, $f_\star = 0$ for readability):

Insight into proof structures?

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable) with $f \in \mathcal{F}$. Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x).$$

For example, assume ($x_\star = 0$, $f_\star = 0$ for readability):

- ◇ Gradient-based method, e.g.: $x_k = x_{k-1} - \sum_{i=0}^{k-1} h_{k,i} \nabla f(x_i)$,

Insight into proof structures?

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable) with $f \in \mathcal{F}$. Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x).$$

For example, assume ($x_\star = 0$, $f_\star = 0$ for readability):

- ◇ Gradient-based method, e.g.: $x_k = x_{k-1} - \sum_{i=0}^{k-1} h_{k,i} \nabla f(x_i)$,
- ◇ $\mathcal{F}$ can be described through nice linear/quadratic (interpolation) inequalities

Insight into proof structures?

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable) with $f \in \mathcal{F}$. Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x).$$

For example, assume ($x_\star = 0$, $f_\star = 0$ for readability):

- Gradient-based method, e.g.: $x_k = x_{k-1} - \sum_{i=0}^{k-1} h_{k,i} \nabla f(x_i)$,
- $\mathcal{F}$ can be described through nice linear/quadratic (interpolation) inequalities in terms of $F = [f_0, \ldots, f_n]$ and $P = [x_0, \nabla f(x_0), \ldots, \nabla f(x_n)]$ of the form $\mathrm{Tr}(P^T P A_i) + b_i^T F \leqslant 0$.

Insight into proof structures?

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable) with $f \in \mathcal{F}$. Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x).$$

For example, assume ($x_\star = 0$, $f_\star = 0$ for readability):

- Gradient-based method, e.g.: $x_k = x_{k-1} - \sum_{i=0}^{k-1} h_{k,i} \nabla f(x_i)$,
- $\mathcal{F}$ can be described through nice linear/quadratic (interpolation) inequalities in terms of $F = [f_0, \dots, f_n]$ and $P = [x_0, \nabla f(x_0), \dots, \nabla f(x_n)]$ of the form $\mathrm{Tr}(P^T P A_i) + b_i^T F \leqslant 0$.
- Performance criterion/initial condition linear in $(G = P^T P, F)$.

Insight into proof structures?

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable) with $f \in \mathcal{F}$. Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x).$$

For example, assume ($x_\star = 0$, $f_\star = 0$ for readability):

- ◇ Gradient-based method, e.g.: $x_k = x_{k-1} - \sum_{i=0}^{k-1} h_{k,i} \nabla f(x_i)$,
- ◇ $\mathcal{F}$ can be described through nice linear/quadratic (interpolation) inequalities in terms of $F = [f_0, \ldots, f_n]$ and $P = [x_0, \nabla f(x_0), \ldots, \nabla f(x_n)]$ of the form $\mathrm{Tr}(P^T P A_i) + b_i^T F \leqslant 0$.
- ◇ Performance criterion/initial condition linear in $(G = P^T P, F)$.

Then, proofs of the fact "Perf $\leqslant \tau$Init"

Insight into proof structures?

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable) with $f \in \mathcal{F}$. Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x).$$

For example, assume ($x_\star = 0$, $f_\star = 0$ for readability):

- ◇ Gradient-based method, e.g.: $x_k = x_{k-1} - \sum_{i=0}^{k-1} h_{k,i} \nabla f(x_i)$,
- ◇ $\mathcal{F}$ can be described through nice linear/quadratic (interpolation) inequalities in terms of $F = [f_0, \ldots, f_n]$ and $P = [x_0, \nabla f(x_0), \ldots, \nabla f(x_n)]$ of the form $\mathrm{Tr}(P^T P A_i) + b_i^T F \leqslant 0$.
- ◇ Performance criterion/initial condition linear in $(G = P^T P, F)$.

Then, proofs of the fact "Perf $\leqslant \tau$Init" can be framed as: $\exists (\lambda_i)_i$, $\lambda_i \geqslant 0$, and $S \succcurlyeq 0$

$$0 \geqslant \sum_i \lambda_i \left(\mathrm{Tr}(G A_i) + b_i^T F \right) = \text{Perf} - \tau \text{Init} + \mathrm{Tr}(GS).$$

Back to legitimate questions

Legitimate questions (gradient descent, one iteration):

- ◇ anything improvable? Realistic analyses?
- ◇ How to choose the right inequalities to combine?
- ◇ Why studying this specific quantity? Possible to adapt to other quantities?
- ◇ Unique way to arrive to the desired result?
- ◇ How likely are we to find such proofs in more complicated cases?

Software: PEPit/PESTO

PEPit

Search docs

CONTENTS:

- Welcome to PEPit's documentation!
- Quick start guide
- API and modules
- Examples
- What's new in PEPit
- Contributing

Welcome to PEPit's documentation! — View page source

Welcome to PEPit's documentation!

Contents:

- Welcome to PEPit's documentation!
- Quick start guide
- API and modules
- Examples
- What's new in PEPit
- Contributing

PEPit: Performance Estimation in Python

Tests passing | codecov 89% | docs passing | pypi package 0.3.2 | downloads 29k | license MIT

This open source Python library provides a generic way to use PEP framework in Python. Performance estimation problems were introduced in 2014 by **Yoel Drori** and **Marc Teboulle**, see [1]. PEPit is mainly based on the formalism and developments from [2, 3] by a subset of the authors of this toolbox. A friendly informal introduction to this formalism is available in this blog post and a corresponding Matlab library is presented in [4] (PESTO).

Website and documentation of PEPit: https://pepit.readthedocs.io/

Source Code (MIT): https://github.com/PerformanceEstimation/PEPit

Using and citing the toolbox

This code comes jointly with the following `reference`:

```
B. Goujaud, C. Moucer, F. Glineur, J. Hendrickx, A. Taylor, A. Dieuleveut (2022).
```

Examples: gradient methods and momentum

$$\min_{x \in \mathbb{R}^d} f(x),$$

with f an L-smooth convex function.

Examples: gradient methods and momentum

$$\min_{x \in \mathbb{R}^d} f(x),$$

with f an L-smooth convex function. Compare three algorithms:

Examples: gradient methods and momentum

$$\min_{x \in \mathbb{R}^d} f(x),$$

with f an L-smooth convex function. Compare three algorithms:

- Gradient descent: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$.

Examples: gradient methods and momentum

$$\min_{x \in \mathbb{R}^d} f(x),$$

with f an L-smooth convex function. Compare three algorithms:

- ◇ Gradient descent: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$.
- ◇ Heavy-ball method $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$

Examples: gradient methods and momentum

$$\min_{x \in \mathbb{R}^d} f(x),$$

with f an L-smooth convex function. Compare three algorithms:

- Gradient descent: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$.
- Heavy-ball method $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$
  (choice: $\alpha = \frac{1}{2L}$, $\beta = \sqrt{1 - L\alpha}$)

Examples: gradient methods and momentum

$$\min_{x \in \mathbb{R}^d} f(x),$$

with f an L-smooth convex function. Compare three algorithms:

- Gradient descent: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$.
- Heavy-ball method $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$
  (choice: $\alpha = \frac{1}{2L}$, $\beta = \sqrt{1 - L\alpha}$)
- Accelerated gradient method:

$$x_{k+1} = y_k - \tfrac{1}{L}\nabla f(x_k)$$
$$y_{k+1} = x_{k+1} + \tfrac{k-1}{k+1}(x_{k+1} - x_k).$$

Examples: gradient methods and momentum

$$\min_{x \in \mathbb{R}^d} f(x),$$

with f an L-smooth convex function. Compare three algorithms:

- Gradient descent: $x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$.
- Heavy-ball method $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$
  (choice: $\alpha = \frac{1}{2L}$, $\beta = \sqrt{1 - L\alpha}$)
- Accelerated gradient method:

$$x_{k+1} = y_k - \tfrac{1}{L}\nabla f(x_k)$$
$$y_{k+1} = x_{k+1} + \tfrac{k-1}{k+1}(x_{k+1} - x_k).$$

What can we guarantee on...

$$\frac{f(x_n) - f(x_\star)}{\|x_0 - x_\star\|^2} \leqslant ? \qquad \frac{\|\nabla f(x_n)\|^2}{\|x_0 - x_\star\|^2} \leqslant ? \qquad \frac{\min\limits_{0 \leqslant k \leqslant n} \|\nabla f(x_k)\|^2}{\|x_0 - x_\star\|^2} \leqslant ?$$

Example: a primal-dual proximal point

Minimize sum of two convex (ccp) functions

$$\min_{x \in \mathbb{R}^d} f(x) + h(x)$$

assume $\exists x_\star, y_\star$ (KKT point): $-y_\star \in \partial f(x_\star)$, $x_\star \in \partial h^*(y_\star)$.

Example: a primal-dual proximal point

Minimize sum of two convex (ccp) functions

$$\min_{x \in \mathbb{R}^d} f(x) + h(x)$$

assume $\exists x_\star, y_\star$ (KKT point): $-y_\star \in \partial f(x_\star)$, $x_\star \in \partial h^*(y_\star)$.

Primal-dual proximal point algorithm (see, e.g., [Rockafellar, 1976])

Input: f, h convex (ccp) functions, $(y_0, x_0) \in \mathbb{R}^d \times \mathbb{R}^d$.

For $k = 0, 1, \ldots$

$$(y_{k+1}, x_{k+1}) = \operatorname*{argmax}_{y \in \mathbb{R}^d} \operatorname*{argmin}_{x \in \mathbb{R}^d} \Big\{ f(x) - h^*(y) + \langle y, x \rangle + \tfrac{1}{2\alpha}\|x - x_k\|^2 - \tfrac{1}{2\alpha}\|y - y_k\|^2 \Big\}$$

Example: a primal-dual proximal point

Minimize sum of two convex (ccp) functions

$$\min_{x\in\mathbb{R}^d} f(x) + h(x)$$

assume $\exists x_\star, y_\star$ (KKT point): $-y_\star \in \partial f(x_\star)$, $x_\star \in \partial h^*(y_\star)$.

Primal-dual proximal point algorithm (see, e.g., [Rockafellar, 1976])

Input: f, h convex (ccp) functions, $(y_0, x_0) \in \mathbb{R}^d \times \mathbb{R}^d$.

For $k = 0, 1, \ldots$

$$(y_{k+1}, x_{k+1}) = \operatorname*{argmax}_{y\in\mathbb{R}^d} \operatorname*{argmin}_{x\in\mathbb{R}^d} \Big\{ f(x) - h^*(y) + \langle y,\, x\rangle + \tfrac{1}{2\alpha}\|x - x_k\|^2 - \tfrac{1}{2\alpha}\|y - y_k\|^2 \Big\}$$

What guarantees of type (for some elements of ∂f and ∂h^*)

$$\frac{\|\partial f(x_n) + y_n\|^2 + \|x_n - \partial h^*(y_n)\|^2}{\|x_0 - x_\star\|^2 + \|y_0 - y_\star\|^2} \leqslant \tau(n, \alpha)?$$

Recap'

Recap'

- ☺ Worst-case guarantees *cannot be improved*, systematic approach,
- ☺ allows reaching analyses that could barely be obtained by hand,
- ☺ fair amount of scenarios/algorithms (e.g., proximal terms, stochastic, etc.),
- ☹ SDPs typically become prohibitively large in a variety of scenarios,
- ☹ transient behavior VS. asymptotic behavior: might be hard to distinguish with small N,
- ☹ proofs (may be) quite involved and hard to intuit,
- ☹ proofs (may be) hard to generalize.

A few research directions

Directions & difficulties to go further?

- ◇ characterization of problem classes (interpolation),
  (account for problem structures)
- ◇ structured proofs,
- ◇ algorithm design,
- ◇ symbolic solves & formalization,
- ◇ packages for ease of use.

Constructive approach to performance analysis

Structured analyses

Towards optimal algorithms

Concluding remarks

Constructive approach to performance analysis

Structured analyses

Towards optimal algorithms

Concluding remarks

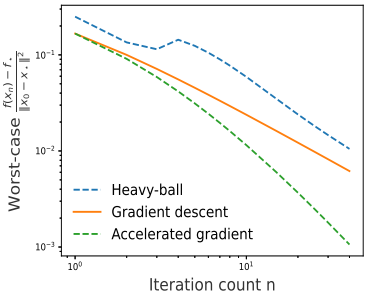
Structured performance analyses

Structured performance analyses

So far: we searched for iteration-dependent analyses.

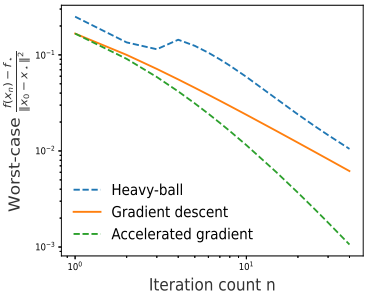
Structured performance analyses

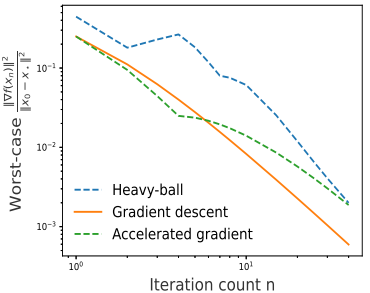
So far: we searched for iteration-dependent analyses.



Structured performance analyses

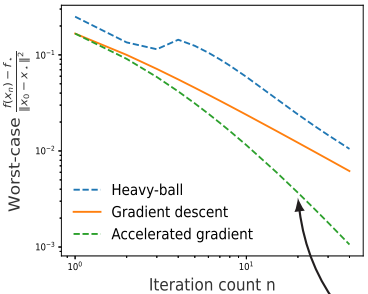
So far: we searched for iteration-dependent analyses.

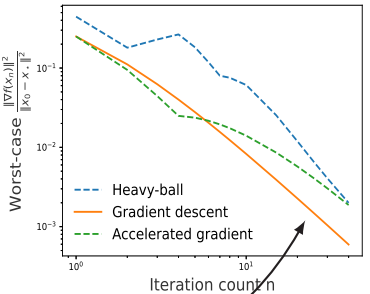




Structured performance analyses

So far: we searched for iteration-dependent analyses.





What to expect for larger n?

Structured proofs

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable). Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x),$$

with f is L-smooth and μ-strongly convex ($f \in \mathcal{F}_{\mu,L}$).

Structured proofs

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable). Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x),$$

with f is L-smooth and μ-strongly convex ($f \in \mathcal{F}_{\mu,L}$).

(Gradient method) We decide to use: $x_{k+1} = x_k - \alpha \nabla f(x_k)$

Structured proofs

Let $f : \mathbb{R}^d \to \mathbb{R}$ (continuously differentiable). Find $x_\star \in \mathbb{R}^d$ such that

$$f(x_\star) = \min_{x \in \mathbb{R}^d} f(x),$$

with f is L-smooth and μ-strongly convex ($f \in \mathcal{F}_{\mu,L}$).

(Gradient method) We decide to use: $x_{k+1} = x_k - \alpha \nabla f(x_k)$

Question: what *a priori* convergence rate?

Examples: we have seen $\frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2}$, what about $\frac{\|\nabla f(x_1)\|^2}{\|\nabla f(x_0)\|^2}$, $\frac{f(x_1) - f_\star}{f(x_0) - f_\star}$?

Structured proofs

Searching for a better rate?

Structured proofs

Searching for a better rate?

- ◇ Any of the above is fine: $\frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2}$, $\frac{\|\nabla f(x_1)\|^2}{\|\nabla f(x_0)\|^2}$, $\frac{f(x_1) - f_\star}{f(x_0) - f_\star}$

Structured proofs

Searching for a better rate?

- ◇ Any of the above is fine: $\frac{\|x_1 - x_\star\|^2}{\|x_0 - x_\star\|^2}$, $\frac{\|\nabla f(x_1)\|^2}{\|\nabla f(x_0)\|^2}$, $\frac{f(x_1) - f_\star}{f(x_0) - f_\star}$ $\rightarrow$ we could pick the *best*.

Structured proofs

Searching for a better rate?

- ◇ Any of the above is fine: $\frac{\|x_1-x_\star\|^2}{\|x_0-x_\star\|^2}$, $\frac{\|\nabla f(x_1)\|^2}{\|\nabla f(x_0)\|^2}$, $\frac{f(x_1)-f_\star}{f(x_0)-f_\star}$ $\rightarrow$ we could pick the *best*.
- ◇ Any combination is fine (e.g., with $a, b \geqslant 0$, $a+b=1$), e.g.,

$$\frac{a\|x_1-x_\star\|^2+b\|\nabla f(x_1)\|^2}{a\|x_0-x_\star\|^2+b\|\nabla f(x_0)\|^2}.$$

Structured proofs

Searching for a better rate?

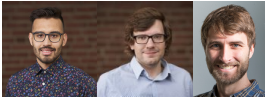
- ◇ Any of the above is fine: $\frac{\|x_1-x_\star\|^2}{\|x_0-x_\star\|^2}$, $\frac{\|\nabla f(x_1)\|^2}{\|\nabla f(x_0)\|^2}$, $\frac{f(x_1)-f_\star}{f(x_0)-f_\star}$ $\rightarrow$ we could pick the *best*.
- ◇ Any combination is fine (e.g., with $a, b \geqslant 0$, $a+b=1$), e.g.,

$$\frac{a\|x_1-x_\star\|^2+b\|\nabla f(x_1)\|^2}{a\|x_0-x_\star\|^2+b\|\nabla f(x_0)\|^2}.$$

- ◇ What about optimizing over it?

$$\min_{\substack{a,b\geqslant 0\\ a+b=1}} \max_{\substack{f\in\mathcal{F}_{\mu,L}\\ x_0,x_1,x_\star}} \left\{ \frac{a\|x_1-x_\star\|^2+b\|\nabla f(x_1)\|^2}{a\|x_0-x_\star\|^2+b\|\nabla f(x_0)\|^2} : \; x_1 = x_0 - \alpha\nabla f(x_0) \right\}.$$

- ◇ ... yet, one can be smarter (ask Pontus & Manu)



Cycles

Heavy-ball

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).$$

Pick specific (α, β) and fix cycle length K.

Cycles

Heavy-ball

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).$$

Pick specific (α, β) and fix cycle length K.

Look for non-trivial cycles of length $K \in \mathbb{N}$ by solving:

$$\min_{\substack{f \\ x_0, \ldots, x_{K+1}}} \quad \|x_K - x_0\|^2 + \|x_{K+1} - x_1\|^2$$

$$\begin{aligned} \text{s.t. } & f \in \mathcal{F}_{\mu,L} && \text{Functional class} \\ & x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) && \text{Algorithm} \\ & \|x_0 - x_1\|^2 \geqslant 1 && \text{Non-trivial cycle} \end{aligned}$$

Cycles

Heavy-ball

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}).$$

Pick specific (α, β) and fix cycle length K.

Look for non-trivial cycles of length $K \in \mathbb{N}$ by solving:

$$\begin{aligned} \min_{\substack{f \\ x_0,\dots,x_{K+1}}} \quad & \|x_K - x_0\|^2 + \|x_{K+1} - x_1\|^2 \\ \text{s.t. } & f \in \mathcal{F}_{\mu,L} && \text{Functional class} \\ & x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}) && \text{Algorithm} \\ & \|x_0 - x_1\|^2 \geqslant 1 && \text{Non-trivial cycle} \end{aligned}$$

From same steps as before $\rightarrow$ SDP formulation $\rightarrow$ LP (via convexity and symmetries).

Heavy-ball method: Lyapunov vs. cycles

Heavy-ball: $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$. Choices of (α, β) for convergence?[2,3,4]

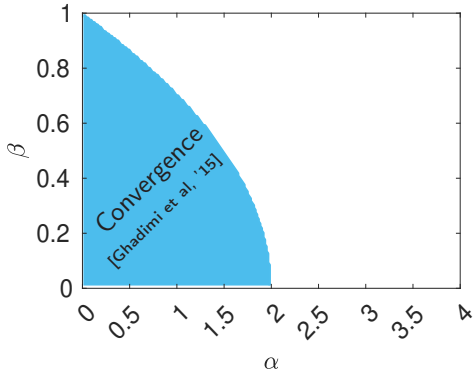
[2] Classical region from [Ghadimi, Feyzmahdavian, Johansson, '15]

[3] Known 3-cycle for optimal quadratic tuning of HB when used beyond quadratics [Lessard, Recht, Packard, '16].

[4] Goujaud, T, Dieuleveut ('25). *Provable non-accelerations of the heavy-ball method.*

Heavy-ball method: Lyapunov vs. cycles

Heavy-ball: $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$. Choices of (α, β) for convergence?[2,3,4]



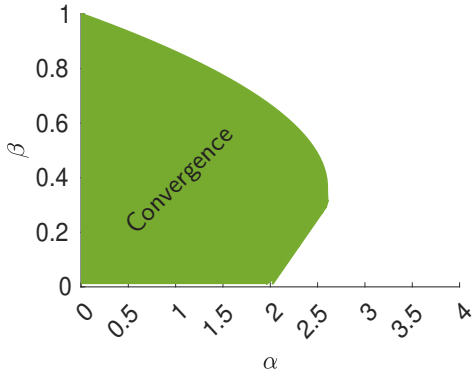
[2] Classical region from [Ghadimi, Feyzmahdavian, Johansson, '15]

[3] Known 3-cycle for optimal quadratic tuning of HB when used beyond quadratics [Lessard, Recht, Packard, '16].

[4] Goujaud, T, Dieuleveut ('25). *Provable non-accelerations of the heavy-ball method.*

Heavy-ball method: Lyapunov vs. cycles

Heavy-ball: $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$. Choices of (α, β) for convergence?[2,3,4]



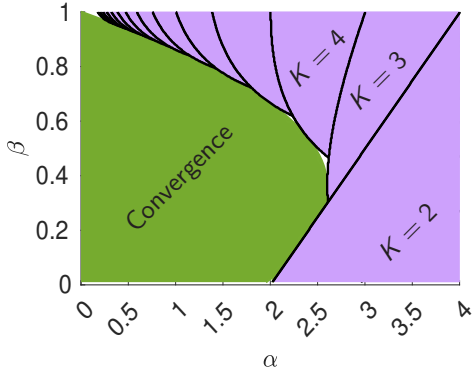
[2] Classical region from [Ghadimi, Feyzmahdavian, Johansson, '15]

[3] Known 3-cycle for optimal quadratic tuning of HB when used beyond quadratics [Lessard, Recht, Packard, '16].

[4] Goujaud, T, Dieuleveut ('25). *Provable non-accelerations of the heavy-ball method.*

Heavy-ball method: Lyapunov vs. cycles

Heavy-ball: $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$. Choices of (α, β) for convergence?[2,3,4]



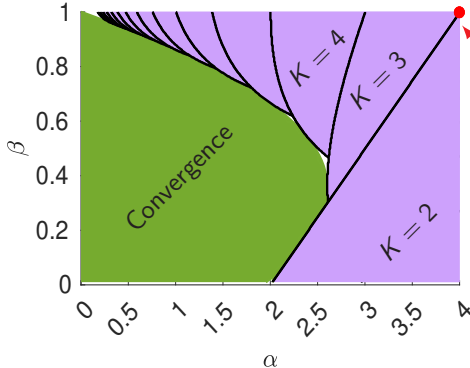
[2] Classical region from [Ghadimi, Feyzmahdavian, Johansson, '15]

[3] Known 3-cycle for optimal quadratic tuning of HB when used beyond quadratics [Lessard, Recht, Packard, '16].

[4] Goujaud, T, Dieuleveut ('25). *Provable non-accelerations of the heavy-ball method.*

Heavy-ball method: Lyapunov vs. cycles

Heavy-ball: $x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1})$. Choices of (α, β) for convergence?[2,3,4]



"Optimal tuning" for quadratic optimization[8]

(numerics for $L/\mu = 10^7$)

[2] Classical region from [Ghadimi, Feyzmahdavian, Johansson, '15]

[3] Known 3-cycle for optimal quadratic tuning of HB when used beyond quadratics [Lessard, Recht, Packard, '16].

[4] Goujaud, T, Dieuleveut ('25). *Provable non-accelerations of the heavy-ball method.*

Constructive approach to performance analysis

Structured analyses

Towards optimal algorithms

Concluding remarks

Designing algorithms

A "generic" first-order method

$$
\begin{aligned}
w_1 &= w_0 - \alpha_{1,0}\nabla f(w_0) \\
w_2 &= w_1 - \alpha_{2,0}\nabla f(w_0) - \alpha_{2,1}\nabla f(w_1) \\
w_3 &= w_2 - \alpha_{3,0}\nabla f(w_0) - \alpha_{3,1}\nabla f(w_1) - \alpha_{3,2}\nabla f(w_2) \\
&\vdots \\
w_n &= w_{n-1} - \sum_{i=0}^{n-1} \alpha_{n,i}\nabla f(w_i),
\end{aligned}
\tag{FOM}
$$

for some coefficients $\{\alpha_{i,j}\}$. Generic **non-adaptive** first-order method.

Designing algorithms

A "generic" first-order method

$$
\begin{aligned}
w_1 &= w_0 - \alpha_{1,0}\nabla f(w_0) \\
w_2 &= w_1 - \alpha_{2,0}\nabla f(w_0) - \alpha_{2,1}\nabla f(w_1) \\
w_3 &= w_2 - \alpha_{3,0}\nabla f(w_0) - \alpha_{3,1}\nabla f(w_1) - \alpha_{3,2}\nabla f(w_2) \\
&\vdots \\
w_n &= w_{n-1} - \sum_{i=0}^{n-1} \alpha_{n,i}\nabla f(w_i),
\end{aligned}
\tag{FOM}
$$

for some coefficients $\{\alpha_{i,j}\}$. Generic **non-adaptive** first-order method.

How to choose $\{\alpha_{i,j}\}$?

Designing algorithms

A "generic" first-order method

$$
\begin{aligned}
w_1 &= w_0 - \alpha_{1,0}\nabla f(w_0) \\
w_2 &= w_1 - \alpha_{2,0}\nabla f(w_0) - \alpha_{2,1}\nabla f(w_1) \\
w_3 &= w_2 - \alpha_{3,0}\nabla f(w_0) - \alpha_{3,1}\nabla f(w_1) - \alpha_{3,2}\nabla f(w_2) \\
&\vdots \\
w_n &= w_{n-1} - \sum_{i=0}^{n-1} \alpha_{n,i}\nabla f(w_i),
\end{aligned}
\tag{FOM}
$$

for some coefficients $\{\alpha_{i,j}\}$. Generic **non-adaptive** first-order method.

How to choose $\{\alpha_{i,j}\}$?

- ◇ pick a performance criterion, for instance $\frac{\|w_n - w_\star\|^2}{\|w_0 - w_\star\|^2}$,

Designing algorithms

A "generic" first-order method

$$
\begin{aligned}
w_1 &= w_0 - \alpha_{1,0}\nabla f(w_0) \\
w_2 &= w_1 - \alpha_{2,0}\nabla f(w_0) - \alpha_{2,1}\nabla f(w_1) \\
w_3 &= w_2 - \alpha_{3,0}\nabla f(w_0) - \alpha_{3,1}\nabla f(w_1) - \alpha_{3,2}\nabla f(w_2) \\
&\vdots \\
w_n &= w_{n-1} - \sum_{i=0}^{n-1} \alpha_{n,i}\nabla f(w_i),
\end{aligned}
\tag{FOM}
$$

for some coefficients $\{\alpha_{i,j}\}$. Generic **non-adaptive** first-order method.

How to choose $\{\alpha_{i,j}\}$?

- ◇ pick a performance criterion, for instance $\frac{\|w_n - w_\star\|^2}{\|w_0 - w_\star\|^2}$,
- ◇ solve the minimax (minimize worst-case): $\min_{\{\alpha_{i,j}\}_{i,j}} \max_{f\in\mathcal{F},\{w_i\}} \frac{\|w_n - w_\star\|^2}{\|w_0 - w_\star\|^2}$.

Traditional design

Algorithm design (with guarantees):

Traditional design

Algorithm design (with guarantees):

- ◇ unconstrained quadratic programming:

  first-order methods $\Leftrightarrow$ polynomials $\rightarrow$ optimal algorithms via optimal polynomials.[5,6,7]

[5] Golub and Varga (1961). "Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods."

[6] Polyak (1987). "Introduction to optimization."

[7] Nemirovski (1995). "Information-based complexity of convex programming." (Lecture notes)

Traditional design

Algorithm design (with guarantees):

- ◇ unconstrained quadratic programming:

  first-order methods $\Leftrightarrow$ polynomials $\rightarrow$ optimal algorithms via optimal polynomials.[5,6,7]

- ◇ Beyond quadratics: traditionally more "hand-crafted"
  - – Analogy with conjugate gradients.[8,9]
  - – Lyapunov function-type analyses.[10,11]

---

[5] Golub and Varga (1961). "Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods."

[6] Polyak (1987). "Introduction to optimization."

[7] Nemirovski (1995). "Information-based complexity of convex programming." (Lecture notes)

[8] Nemirovski (1982). "Orth-method for smooth convex optimization."

[9] Narkiss and Zibulevsky (2005). "Sequential subspace optimization method for large-scale unconstrained problems."

[10] Nesterov (1983). "A method for solving the convex programming problem with convergence rate $O(1/k^2)$."

[11] Beck and Teboulle ('09). "A fast iterative shrinkage-thresholding algorithm for linear inverse problems."

Design problem

How to solve the design problem (or proxy of it)?

$$\min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{\|w_n - w_\star\|^2}{\|w_0 - w_\star\|^2} ?$$

Design problem

How to solve the design problem (or proxy of it)?

$$\min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{\|w_n - w_\star\|^2}{\|w_0 - w_\star\|^2}? \qquad \min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{f(w_n) - f(w_\star)}{f(w_0) - f(w_\star)}?$$

Design problem

How to solve the design problem (or proxy of it)?

$$\min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{\|w_n - w_\star\|^2}{\|w_0 - w_\star\|^2}? \quad \min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{f(w_n) - f(w_\star)}{f(w_0) - f(w_\star)}? \quad \min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{\|\nabla f(w_n)\|^2}{\|\nabla f(w_0)\|^2}?$$

Design problem

How to solve the design problem (or proxy of it)?

$$\min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{\|w_n - w_\star\|^2}{\|w_0 - w_\star\|^2}? \quad \min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{f(w_n) - f(w_\star)}{f(w_0) - f(w_\star)}? \quad \min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{\|\nabla f(w_n)\|^2}{\|\nabla f(w_0)\|^2}?$$

- Convex relaxations,

Design problem

How to solve the design problem (or proxy of it)?

$$\min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{\|w_n - w_\star\|^2}{\|w_0 - w_\star\|^2}? \quad \min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{f(w_n) - f(w_\star)}{f(w_0) - f(w_\star)}? \quad \min_{\{\alpha_{i,j}\}} \max_{f \in \mathcal{F}} \frac{\|\nabla f(w_n)\|^2}{\|\nabla f(w_0)\|^2}?$$

- ◇ Convex relaxations,
- ◇ analogies (e.g., with conjugate gradient methods)

$$w_{k+1} \in \operatorname*{argmin}_{x \in \mathbb{R}^d} \{f(w) : \ w \in w_0 + \operatorname{span}\{\nabla f(w_0), \ldots, \nabla f(w_k)\}\},$$

Design problem

How to solve the design problem (or proxy of it)?

$$\min_{\{\alpha_{i,j}\}} \max_{f\in\mathcal{F}} \frac{\|w_n - w_\star\|^2}{\|w_0 - w_\star\|^2}? \quad \min_{\{\alpha_{i,j}\}} \max_{f\in\mathcal{F}} \frac{f(w_n) - f(w_\star)}{f(w_0) - f(w_\star)}? \quad \min_{\{\alpha_{i,j}\}} \max_{f\in\mathcal{F}} \frac{\|\nabla f(w_n)\|^2}{\|\nabla f(w_0)\|^2}?$$

- ◇ Convex relaxations,
- ◇ analogies (e.g., with conjugate gradient methods)

$$w_{k+1} \in \operatorname*{argmin}_{x\in\mathbb{R}^d} \{f(w) : \ w \in w_0 + \operatorname{span}\{\nabla f(w_0), \ldots, \nabla f(w_k)\}\},$$

- ◇ brutal approaches.

Fixed stepsize policy from WC analysis of linesearch?

Target convergence rate (example):

Fixed stepsize policy from WC analysis of linesearch?

Target convergence rate (example):

$$\rho \overset{\text{(def)}}{=} \max_{w_0, w_1, f \in \mathcal{F}_{\mu,\mathrm{L}}} \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \langle \nabla f(w_1), \nabla f(w_0) \rangle = 0, \ \langle \nabla f(w_1), w_1 - w_0 \rangle = 0,$$

Fixed stepsize policy from WC analysis of linesearch?

Target convergence rate (example):

$$\rho \overset{\text{(def)}}{=} \max_{w_0, w_1, f \in \mathcal{F}_{\mu,\mathrm{L}}} \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \langle \nabla f(w_1), \nabla f(w_0) \rangle = 0, \ \langle \nabla f(w_1), w_1 - w_0 \rangle = 0,$$

upper bound from a Lagrangian relaxation with $\lambda_1, \lambda_2 \in \mathbb{R}$:

$$\rho \leqslant \bar{\rho}(\lambda_1, \lambda_2) \overset{\text{(def)}}{=} \max_{x_0, x_1, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} + \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle \right\}.$$

Fixed stepsize policy from WC analysis of linesearch?

Target convergence rate (example):

$$\rho \overset{\text{(def)}}{=} \max_{w_0, w_1, f \in \mathcal{F}_{\mu,L}} \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \langle \nabla f(w_1), \nabla f(w_0) \rangle = 0, \ \langle \nabla f(w_1), w_1 - w_0 \rangle = 0,$$

upper bound from a Lagrangian relaxation with $\lambda_1, \lambda_2 \in \mathbb{R}$:

$$\rho \leqslant \bar{\rho}(\lambda_1, \lambda_2) \overset{\text{(def)}}{=} \max_{x_0, x_1, f \in \mathcal{F}_{\mu,L}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} + \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle \right\}.$$

We can also create an intermediary problem

$$\rho \leqslant \qquad\qquad \leqslant \bar{\rho}(\lambda_1, \lambda_2).$$

Fixed stepsize policy from WC analysis of linesearch?

Target convergence rate (example):

$$\rho \overset{\text{(def)}}{=} \max_{w_{\mathbf{0}}, w_{\mathbf{1}}, f \in \mathcal{F}_{\mu,\mathrm{L}}} \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \langle \nabla f(w_1), \nabla f(w_0) \rangle = 0, \ \langle \nabla f(w_1), w_1 - w_0 \rangle = 0,$$

upper bound from a Lagrangian relaxation with $\lambda_1, \lambda_2 \in \mathbb{R}$:

$$\rho \leqslant \bar{\rho}(\lambda_1, \lambda_2) \overset{\text{(def)}}{=} \max_{x_{\mathbf{0}}, x_{\mathbf{1}}, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} + \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle \right\}.$$

We can also create an intermediary problem

$$\rho \leqslant \max_{w_{\mathbf{0}}, w_{\mathbf{1}}, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle = 0 \right\} \leqslant \bar{\rho}(\lambda_1, \lambda_2).$$

Fixed stepsize policy from WC analysis of linesearch?

Target convergence rate (example):

$$\rho \overset{\text{(def)}}{=} \max_{w_{\mathbf{0}}, w_{\mathbf{1}}, f \in \mathcal{F}_{\mu,\mathrm{L}}} \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \langle \nabla f(w_1), \nabla f(w_0) \rangle = 0, \ \langle \nabla f(w_1), w_1 - w_0 \rangle = 0,$$

upper bound from a Lagrangian relaxation with $\lambda_1, \lambda_2 \in \mathbb{R}$:

$$\rho \leqslant \bar{\rho}(\lambda_1, \lambda_2) \overset{\text{(def)}}{=} \max_{x_{\mathbf{0}}, x_{\mathbf{1}}, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} + \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle \right\}.$$

We can also create an intermediary problem

$$\rho \leqslant \max_{w_{\mathbf{0}}, w_{\mathbf{1}}, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle = 0 \right\} \leqslant \bar{\rho}(\lambda_1, \lambda_2).$$

So, for any pair $\lambda_1, \lambda_2 \in \mathbb{R}$ we get:

Fixed stepsize policy from WC analysis of linesearch?

Target convergence rate (example):

$$\rho \overset{\text{(def)}}{=} \max_{w_{\mathbf{0}}, w_{\mathbf{1}}, f \in \mathcal{F}_{\mu,\mathrm{L}}} \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \langle \nabla f(w_1), \nabla f(w_0) \rangle = 0, \ \langle \nabla f(w_1), w_1 - w_0 \rangle = 0,$$

upper bound from a Lagrangian relaxation with $\lambda_1, \lambda_2 \in \mathbb{R}$:

$$\rho \leqslant \bar{\rho}(\lambda_1, \lambda_2) \overset{\text{(def)}}{=} \max_{x_{\mathbf{0}}, x_{\mathbf{1}}, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} + \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle \right\}.$$

We can also create an intermediary problem

$$\rho \leqslant \max_{w_{\mathbf{0}}, w_{\mathbf{1}}, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle = 0 \right\} \leqslant \bar{\rho}(\lambda_1, \lambda_2).$$

So, for any pair $\lambda_1, \lambda_2 \in \mathbb{R}$ we get:

- an upper bound $\bar{\rho}(\lambda_1, \lambda_2)$ (possibly $+\infty$) on ρ,

Fixed stepsize policy from WC analysis of linesearch?

Target convergence rate (example):

$$\rho \overset{\text{(def)}}{=} \max_{w_0, w_1, f \in \mathcal{F}_{\mu,\mathrm{L}}} \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \langle \nabla f(w_1), \nabla f(w_0) \rangle = 0, \ \langle \nabla f(w_1), w_1 - w_0 \rangle = 0,$$

upper bound from a Lagrangian relaxation with $\lambda_1, \lambda_2 \in \mathbb{R}$:

$$\rho \leqslant \bar{\rho}(\lambda_1, \lambda_2) \overset{\text{(def)}}{=} \max_{x_0, x_1, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} + \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle \right\}.$$

We can also create an intermediary problem

$$\rho \leqslant \max_{w_0, w_1, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle = 0 \right\} \leqslant \bar{\rho}(\lambda_1, \lambda_2).$$

So, for any pair $\lambda_1, \lambda_2 \in \mathbb{R}$ we get:

- an upper bound $\bar{\rho}(\lambda_1, \lambda_2)$ (possibly $+\infty$) on ρ,
- all methods satisfying $\langle \nabla f(w_1), \lambda_1 \nabla f(w_0) + \lambda_2 (w_1 - w_0) \rangle = 0$ have convergence rate at most $\bar{\rho}(\lambda_1, \lambda_2)$.

Fixed stepsize policy from WC analysis of linesearch?

Target convergence rate (example):

$$\rho \overset{\text{(def)}}{=} \max_{w_0, w_1, f \in \mathcal{F}_{\mu,\mathrm{L}}} \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \langle \nabla f(w_1), \nabla f(w_0) \rangle = 0, \ \langle \nabla f(w_1), w_1 - w_0 \rangle = 0,$$

upper bound from a Lagrangian relaxation with $\lambda_1, \lambda_2 \in \mathbb{R}$:

$$\rho \leqslant \bar{\rho}(\lambda_1, \lambda_2) \overset{\text{(def)}}{=} \max_{x_0, x_1, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} + \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle \right\}.$$

We can also create an intermediary problem

$$\rho \leqslant \max_{w_0, w_1, f \in \mathcal{F}_{\mu,\mathrm{L}}} \left\{ \frac{f(w_1) - f_\star}{f(w_0) - f_\star} \text{ s.t. } \lambda_1 \langle \nabla f(w_1), \nabla f(w_0) \rangle + \lambda_2 \langle \nabla f(w_1), w_1 - w_0 \rangle = 0 \right\} \leqslant \bar{\rho}(\lambda_1, \lambda_2).$$

So, for any pair $\lambda_1, \lambda_2 \in \mathbb{R}$ we get:

- an upper bound $\bar{\rho}(\lambda_1, \lambda_2)$ (possibly $+\infty$) on ρ,
- all methods satisfying $\langle \nabla f(w_1), \lambda_1 \nabla f(w_0) + \lambda_2 (w_1 - w_0) \rangle = 0$ have convergence rate at most $\bar{\rho}(\lambda_1, \lambda_2)$.

Bonus: there exists a choice $\lambda_1^\star, \lambda_2^\star$ such that

$$\rho = \bar{\rho}(\lambda_1^\star, \lambda_2^\star).$$

Design procedure

Algorithm constraints, and associated Lagrange multipliers:

$$\langle \nabla f(w_i), \nabla f(w_j) \rangle = 0, \qquad \text{for all } 0 \leqslant j < i = 1, \dots n \qquad : \beta_{i,j},$$
$$\langle \nabla f(w_i), w_j - w_0 \rangle = 0, \qquad \text{for all } 1 \leqslant j \leqslant i = 1, \dots n \qquad : \gamma_{i,j}.$$

Design procedure

Algorithm constraints, and associated Lagrange multipliers:

$$\langle \nabla f(w_i), \nabla f(w_j) \rangle = 0, \qquad \text{for all } 0 \leqslant j < i = 1, \dots n \qquad : \beta_{i,j},$$
$$\langle \nabla f(w_i), w_j - w_0 \rangle = 0, \qquad \text{for all } 1 \leqslant j \leqslant i = 1, \dots n \qquad : \gamma_{i,j}.$$

... replaced by:

$$\langle \nabla f(w_i), \sum_{j=1}^{i} \beta_{i,j}(w_j - w_0) + \sum_{j=0}^{i-1} \gamma_{i,j} \nabla f(w_j) \rangle = 0 \text{ for all } i = 1, \dots, n$$

Design procedure

Algorithm constraints, and associated Lagrange multipliers:

$$\langle \nabla f(w_i), \nabla f(w_j) \rangle = 0, \qquad \text{for all } 0 \leqslant j < i = 1, \dots n \qquad : \beta_{i,j},$$
$$\langle \nabla f(w_i), w_j - w_0 \rangle = 0, \qquad \text{for all } 1 \leqslant j \leqslant i = 1, \dots n \qquad : \gamma_{i,j}.$$

... replaced by:

$$\langle \nabla f(w_i), \sum_{j=1}^{i} \beta_{i,j}(w_j - w_0) + \sum_{j=0}^{i-1} \gamma_{i,j} \nabla f(w_j) \rangle = 0 \text{ for all } i = 1, \dots, n$$

Subspace-search elimination (abstract)

Design procedure

Algorithm constraints, and associated Lagrange multipliers:

$$\langle \nabla f(w_i), \nabla f(w_j) \rangle = 0, \qquad \text{for all } 0 \leqslant j < i = 1, \dots n \qquad : \beta_{i,j},$$
$$\langle \nabla f(w_i), w_j - w_0 \rangle = 0, \qquad \text{for all } 1 \leqslant j \leqslant i = 1, \dots n \qquad : \gamma_{i,j}.$$

... replaced by:

$$\langle \nabla f(w_i), \sum_{j=1}^{i} \beta_{i,j}(w_j - w_0) + \sum_{j=0}^{i-1} \gamma_{i,j} \nabla f(w_j) \rangle = 0 \text{ for all } i = 1, \dots, n$$

Subspace-search elimination (abstract)

1. Choose $n \geqslant 0$, $\mathcal{F}$.

Design procedure

Algorithm constraints, and associated Lagrange multipliers:

$$\langle \nabla f(w_i), \nabla f(w_j) \rangle = 0, \qquad \text{for all } 0 \leqslant j < i = 1, \dots n \qquad : \beta_{i,j},$$
$$\langle \nabla f(w_i), w_j - w_0 \rangle = 0, \qquad \text{for all } 1 \leqslant j \leqslant i = 1, \dots n \qquad : \gamma_{i,j}.$$

... replaced by:

$$\langle \nabla f(w_i), \sum_{j=1}^{i} \beta_{i,j}(w_j - w_0) + \sum_{j=0}^{i-1} \gamma_{i,j} \nabla f(w_j) \rangle = 0 \text{ for all } i = 1, \dots, n$$

Subspace-search elimination (abstract)

1. Choose $n \geqslant 0$, $\mathcal{F}$.
2. Find a feasible solution $\{\beta_{i,j}\}, \{\gamma_{i,j}\}$ to (dual) PEP for the greedy method with rate $\bar{\rho}$.

Design procedure

Algorithm constraints, and associated Lagrange multipliers:

$$\langle \nabla f(w_i), \nabla f(w_j) \rangle = 0, \qquad \text{for all } 0 \leqslant j < i = 1, \dots n \qquad : \beta_{i,j},$$
$$\langle \nabla f(w_i), w_j - w_0 \rangle = 0, \qquad \text{for all } 1 \leqslant j \leqslant i = 1, \dots n \qquad : \gamma_{i,j}.$$

... replaced by:

$$\langle \nabla f(w_i), \sum_{j=1}^{i} \beta_{i,j}(w_j - w_0) + \sum_{j=0}^{i-1} \gamma_{i,j} \nabla f(w_j) \rangle = 0 \text{ for all } i = 1, \dots, n$$

Subspace-search elimination (abstract)

1. Choose $n \geqslant 0$, $\mathcal{F}$.
2. Find a feasible solution $\{\beta_{i,j}\}, \{\gamma_{i,j}\}$ to (dual) PEP for the greedy method with rate $\bar{\rho}$.
3. Any method satisfying

$$\langle \nabla f(w_i), \sum_{j=1}^{i} \beta_{i,j}(w_j - w_0) + \sum_{j=0}^{i-1} \gamma_{i,j} \nabla f(w_j) \rangle = 0 \text{ for all } i = 1, \dots n$$

benefits from the same worst-case convergence rate $\bar{\rho}$.

Optimized gradient methods

Smooth convex minimization setting:

$$\min_{x \in \mathbb{R}^d} f(x)$$

with f being L-smooth and convex.

Optimized gradient methods

Smooth convex minimization setting:

$$\min_{x \in \mathbb{R}^d} f(x)$$

with f being L-smooth and convex.

Lower bound for large-scale setting ($d \geqslant n+2$) by Drori ('17):

$$f(x_n) - f(x_\star) \geqslant \frac{L\|x_0 - x_\star\|^2}{2\theta_n^2},$$

with $\theta_0 = 1$, and:

$$\theta_{i+1} = \begin{cases} \frac{1+\sqrt{4\theta_i^2+1}}{2} & \text{if } i \leqslant n-2, \\ \frac{1+\sqrt{8\theta_i^2+1}}{2} & \text{if } i = n-1. \end{cases}$$

Greedy First-order Method (GFOM)

Inputs: f, x_0, n.

For $i = 1, 2, \ldots, n$

$$x_i = \operatorname*{argmin}_{x \in \mathbb{R}^d} \{f(x) : \; x \in x_0 + \operatorname{span}\{\nabla f(x_0), \ldots, \nabla f(x_{i-1})\}\}.$$

Worst-case guarantee:

$$f(x_n) - f(x_\star) \leqslant \frac{L\|x_0 - x_\star\|^2}{2\theta_n^2}.$$

Optimized gradient methods

As a result from subspace-search elimination, all methods satisfying (for $i = 1, \dots, n$)

$$\langle \nabla f(x_i); x_i - \left[\left(1 - \frac{1}{\theta_i}\right) \left(x_{i-1} - \tfrac{1}{L} \nabla f(x_{i-1})\right) + \frac{1}{\theta_i} \left(x_0 - \tfrac{2}{L} \sum_{j=0}^{i-1} \theta_j \nabla f(x_j) \right) \right] \rangle \leqslant 0$$

benefit from the same guarantee:

$$f(x_n) - f(x_\star) \leqslant \frac{L \|x_0 - x_\star\|^2}{2\theta_n^2},$$

Optimized gradient methods

Three methods with the same (optimal) worst-case behavior

Optimized gradient method with exact line-search

Inputs: f, x_0, n.

For $i = 1, \dots, n$

$$y_i = \left(1 - \frac{1}{\theta_i}\right) x_{i-1} + \frac{1}{\theta_i} x_0$$

$$d_i = \left(1 - \frac{1}{\theta_i}\right) \nabla f(x_{i-1}) + \frac{1}{\theta_i} \left(2 \sum_{j=0}^{i-1} \theta_j \nabla f(x_j)\right)$$

$$\alpha = \operatorname*{argmin}_{\alpha \in \mathbb{R}} f(y_i + \alpha d_i)$$

$$x_i = y_i + \alpha d_i$$

Worst-case guarantee: $f(x_n) - f(x_\star) \leqslant \frac{L\|x_0 - x_\star\|^2}{2\theta_n^2}$

Optimized gradient methods

Three methods with the same (optimal) worst-case behavior

Optimized gradient method

Inputs: f, x_0, n.

For $i = 1, \dots, n$

$$y_i = x_{i-1} - \frac{1}{L}\nabla f(x_{i-1})$$

$$z_i = x_0 - \frac{2}{L}\sum_{j=0}^{i-1} \theta_j \nabla f(x_j)$$

$$x_i = \left(1 - \frac{1}{\theta_i}\right) y_i + \frac{1}{\theta_i} z_i$$

Worst-case guarantee: $f(x_n) - f(x_\star) \leqslant \frac{L\|x_0 - x_\star\|^2}{2\theta_n^2}$.

Constructive approach to performance analysis

Structured analyses

Towards optimal algorithms

Concluding remarks

Concluding remarks

Performance estimation's philosophy

Concluding remarks

Performance estimation's philosophy

- ◇ numerically allows obtaining tight bounds (rigorous baselines),
  - – fast prototyping
  - – worth checking before trying to prove a method works.

Concluding remarks

Performance estimation's philosophy

- ◇ numerically allows obtaining tight bounds (rigorous baselines),
  - – fast prototyping
  - – worth checking before trying to prove a method works.
- ◇ algebraic insights into performance analyses: principled approach,
  - – analyses are dual feasible points,
  - – analyses are linear combinations of certain specific inequalities.

Concluding remarks

Performance estimation's philosophy

- ◇ numerically allows obtaining tight bounds (rigorous baselines),
  - – fast prototyping
  - – worth checking before trying to prove a method works.
- ◇ algebraic insights into performance analyses: principled approach,
  - – analyses are dual feasible points,
  - – analyses are linear combinations of certain specific inequalities.

Byproducts:

- ◇ computer-assisted design of analyses,
- ◇ computer-assisted design of numerical methods,
- ◇ step towards reproducible theory
  - – validation & benchmark tool for proofs (also for reviews ☺),
  - – complements existing open-source initiatives.

A few instructive examples

Worst-case analysis for fixed-point iterations:

- ◇ Lieder ('21). "On the convergence of the Halpern-iteration."

Analysis of the proximal-point algorithm for monotone inclusions:

- ◇ Gu, Yang ('19). "Optimal nonergodic sublinear convergence rate of the proximal point algorithm for maximal monotone inclusion problems."

Application to nonconvex optimization:

- ◇ Abbaszadehpeivasti, de Klerk, Zamani ('22). "The exact worst-case convergence rate of the gradient method with fixed step lengths for L-smooth functions."
- ◇ Rotaru, Glineur, Patrinos ('25). "Exact worst-case convergence rates of gradient descent: a complete analysis for all constant stepsizes over nonconvex and convex functions."

Applications to distributed optimization:

- ◇ Sundararajan, Van Scoy, Lessard ('20). "Analysis and design of first-order distributed optimization algorithms over time-varying graphs."
- ◇ Colla, Hendrickx ('23). "Automatic performance estimation for decentralized optimization."

A few instructive examples

Design first-order methods via PEPs:

- Drori, Teboulle ('14). "Performance of first-order methods for smooth convex minimization: a novel approach."
- Kim, Fessler ('16). "Optimized methods for smooth convex optimization."
- Van Scoy, Freeman, Lynch ('17). "The fastest known globally convergent first-order method for minimizing strongly convex functions."
- Kim, Fessler ('21). "Optimizing the efficiency of first-order methods for decreasing the gradient of smooth convex functions." .
- Altschuler, Parrilo ('23). "Acceleration by Stepsize Hedging I: Multi-Step Descent and the Silver Stepsize Schedule."
- Jang, Das Gupta, Ryu ('25). "Computer-assisted design of accelerated composite optimization methods: OptISTA."

Generic approaches:

- Das Gupta, Van Parys, Ryu ('24). "Branch-and-bound performance estimation programming: A unified methodology for constructing optimal optimization methods."
- Kamri, Hendrickx, Glineur ('25). "Numerical Design of Optimized First-Order Algorithms."

A few references

Historical reference:

- Drori, Teboulle ('14). "Performance of first-order methods for smooth convex minimization: a novel approach."

Main messages:

- T, Hendrickx, Glineur ('17). "Smooth strongly convex interpolation and exact worst-case performance of first-order methods."

- Drori, T ('20). "Efficient first-order methods for convex minimization: a constructive approach."

- Goujaud, Dieuleveut, T ('23). "On fundamental proof structures in first-order optimization."

- Goujaud, Moucer, Glineur, Hendrickx, T, Dieuleveut ('24). "PEPit: computer-assisted worst-case analyses of first-order optimization methods in Python."

- Goujaud, T, Dieuleveut ('25). "Provable non-accelerations of the heavy-ball method."

- Upadhyaya, Banert, T, Giselsson ('25). "Automated tight Lyapunov analysis for first-order methods."

Going further — shameless advertisement

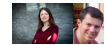
- T, Bach ('19). "Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions."

- Dragomir, T, d'Aspremont, Bolte ('22). "Optimal complexity and certification of Bregman first-order methods."

- T, Drori ('23). "An optimal gradient method for smooth strongly convex minimization."

- Berg Thomsen, T, Dieuleveut ('25). "Tight analyses of first-order methods with error feedback."

- Rubbens, Hendrickx, T ('25). "A constructive approach to strengthen algebraic descriptions of function and operator classes."



- Weibel, Gaillard, Koolen, T ('26). "Optimised projection-free algorithms for online learning: construction and worst-case analysis."

Thanks! Questions?

Great PhD students/visiting PhD students/postdocs at Inria Paris:

Roland Andrews

Daniel Berg Thomsen

Weijia Wang

Si-Yi Meng

Manu Upadhyaya

Fabian Schaipp

Julien Weibel