Greedy Methods for Convex Optimization

From Gradient Descent to Recent Developments

Alex L. Wang, Daniels School of Business, Purdue University

Based on joint work with: **Benjamin Grimmer**, **Kevin Shu**

Outline

- Problem setup: Smooth Convex Optimization
- Gradient Descent (**GD**)
- Fast Gradient Method (**FGM**)
  - Key observation about algorithms and "proofs"
  - Big-O Optimal
- Optimized Gradient Method (**OGM**)
  - Minimax Optimal
- Subgame Perfect Gradient Method (**SPGM**)
  - Subgame Perfect
- If time permits: connections to Acceleration via Stepsize Scheduling

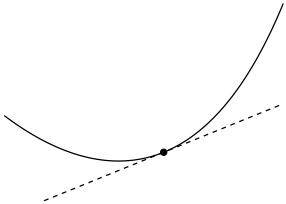
---

[Nesterov 83] [Drori Teboulle 12] [Kim Fessler 16] [Grimmer Shu Wang 23, 24a, 24b] [Altschuler Parrilo 23a, 23b], [Zhang Jiang 24]

Smooth Convex Optimization

Smooth Convex Optimization

- **Task**: Solve $\min_{x \in \mathbb{R}^d} f(x)$
- A priori, we know $f \in \mathcal{F}$:
  - f is 1-smooth: $\|\nabla f(x) - \nabla f(y)\| \leq \|x - y\| \quad \forall x, y$
  - f is convex
  - f has a minimizer
- Black-box first-order oracle setting: $x \mapsto (f(x), \nabla f(x))$



Algorithm Template. N-query first-order method (FOM)

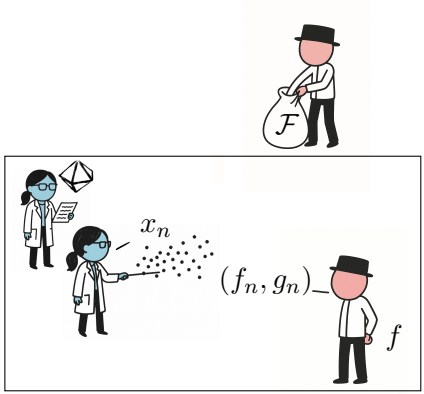
- For $n = 0, \ldots, N$:
  - **Choose** $x_n \in \operatorname{span}\{g_0, \ldots, g_{n-1}\}$
  - **Query first-order oracle** $f_n = f(x_n) \qquad g_n = \nabla f(x_n)$
- Output x_N

[Nemirovski Yudin 83]

The Convex Optimization Game v1

- **Alice** (algorithm), **Bob** (oracle), N number of rounds
- **Bob** secretly chooses $f \in \mathcal{F}$ with $(x_\star, f_\star)$
- Round $n = 0, \ldots, N$
  - **Alice** chooses $x_n \in \operatorname{span}\{g_0, \ldots, g_{n-1}\}$
  - **Bob** returns $f_n = f(x_n)$ and $g_n = \nabla f(x_n)$
- **Alice** pays **Bob** $\frac{f_N - f_\star}{\frac{1}{2}\|x_\star\|^2}$

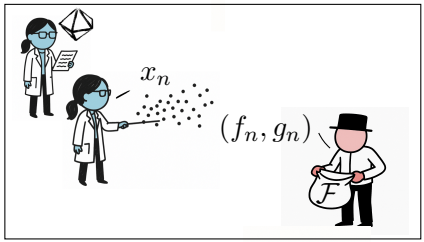
e.g., payoff = $\$\frac{1}{100}$ means
$f(x_N) - f_\star = \frac{1}{100} \cdot \left(\frac{1}{2}\|x_\star\|^2\right)$
Good strategy for **Alice** $\iff$ good FOM



The Convex Optimization Game v2

- **Alice** (algorithm), **Bob** (oracle), N number of rounds
- Round $n = 0, \ldots, N$
  - **Alice** chooses $x_n \in \operatorname{span}\{g_0, \ldots, g_{n-1}\}$
  - **Bob** returns $(f_n, g_n) \in \mathbb{R} \times \mathbb{R}^d$
- **Bob** specifies $f \in \mathcal{F}$ with $(x_\star, f_\star)$

  $f(x_i) = f_i$ and $\nabla f(x_i) = g_i$ for all $i \in [0, N]$
- **Alice** pays **Bob** $\frac{f_N - f_\star}{\frac{1}{2}\|x_\star\|^2}$



Classic FOMs / Strategies for Alice

Gradient Descent (GD)

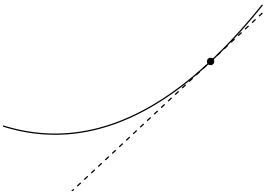
- **Given**: x_{n-1}, g_{n-1}
- **Idea**: Pick $x_n = x_{n-1} - \alpha g_{n-1}$ to maximize the *guaranteed decrease* $f_{n-1} - f_n$

$$\min_{\alpha} \sup_{f \in \mathcal{F}} \left\{ f(x_n) - f_{n-1} : \quad \nabla f(x_{n-1}) = g_{n-1} \right\}$$

- $\alpha = 1$ is the optimal stepsize

$$x_n = x_{n-1}^+ := x_{n-1} - g_{n-1}$$

$$f(x_n) \leq f_{n-1}^+ := f_{n-1} - \frac{1}{2} \|g_{n-1}\|^2$$



Gradient Descent Guarantee

Theorem.

Gradient Descent guarantees: for all $f \in \mathcal{F}$,

$$f(x_N) - f(x_\star) \leq \frac{1}{2N+1}\left(\frac{1}{2}\|x_\star\|^2\right)$$

Alice can guarantee payment $\leq \frac{1}{2N+1}$

- This statement leaves "something on the table" that can be exploited

Understanding Convergence Guarantees and "Proofs"

- **Consider any first-order method**
- At end of iteration n, we have observed first-order history

$$\mathcal{H} = \{(x_0, f_0, g_0), \ldots, (x_n, f_n, g_n)\}$$

Theorem. Key Fact

Suppose $\tau > 0$ and we can prove

$$f_n - f(x_\star) \leq \frac{1}{2\tau} \|x_\star\|^2$$

Then, there exists $z_{n+1}(\mathcal{H})$ so that

$$f_n - f(x_\star) + \frac{1}{2\tau} \|z_{n+1} - x_\star\|^2 \leq \frac{1}{2\tau} \|x_\star\|^2$$

Proof Sketch

The unknown $f \in \mathcal{F}$ must lie in

$$\mathcal{F}_{\mathcal{H}} := \left\{ f \in \mathcal{F} : \begin{array}{ll} f(x_i) = f_i & \forall i \in [0, n] \\ \nabla f(x_i) = g_i & \forall i \in [0, n] \end{array} \right\}$$

Lemma.

The set of possible (optimal value, optimizer)

$$\mathcal{S} := \left\{ (f_\star, x_\star) : \begin{array}{l} \exists f \in \mathcal{F}_{\mathcal{H}} : \\ x_\star \in \arg\min f \\ f(x_\star) = f_\star \end{array} \right\}$$

is polyhedral

$f_n - f_\star \leq \frac{1}{2\tau} \|x_\star\|^2$ states that $\mathcal{S}$ lies above paraboloid

$f_n - f_\star \leq \frac{1}{2\tau} \|x_\star\|^2 - \frac{1}{2\tau} \|z_{n+1} - x_\star\|^2$ states that $\mathcal{S}$ lies above a *linear function*

Fast Gradient Method (FGM)

- **Given**: $(x_{n-1}, \tau_{n-1}, z_n)$ satisfying

$$f_{n-1} - f(x_\star) + \frac{1}{2\tau_{n-1}} \|z_n - x_\star\|^2 \leq \frac{1}{2\tau_{n-1}} \|x_\star\|^2$$

- **Idea**: Two cases
  - $f_{n-1} - f(x_\star) \approx \frac{1}{2\tau_{n-1}} \|x_\star\|^2 \longrightarrow$ set $x_n = z_n$
  - $\|z_n - x_\star\|^2 \approx \|x_\star\|^2 \longrightarrow$ set $x_n = x_{n-1}^+$

  Pick $x_n = (1-\alpha)x_{n-1}^+ + \alpha z_n$ to greedily minimize

$$\min_\alpha \sup_{f \in \mathcal{F}} \left\{ \frac{f(x_n) - f(x_\star)}{\frac{1}{2}\|x_\star\|^2} : f_{n-1} - f(x_\star) + \frac{1}{2\tau_{n-1}} \|z_n - x_\star\|^2 \leq \frac{1}{2\tau_{n-1}} \|x_\star\|^2 \right\}$$

- The optimal choice of $\alpha(\tau_{n-1})$ recovers FGM (see blog post)
- Apply "upgrade" lemma to continue induction

[Nesterov 83]

FGM induction

Lemma.

Suppose (x, τ, z) and $f \in \mathcal{F}$ satisfy **the inductive hypothesis**

$$f(x) - f_\star + \frac{1}{2\tau} \|z - x_\star\|^2 \leq \frac{1}{2\tau} \|x_\star\|^2$$

Then,

$$\hat{x} = \text{some convex combination of } x^+ \text{ and } z$$

$$\hat{\tau} = \tau + \Theta\left(\sqrt{\tau}\right)$$

$$\hat{z} = z - \Theta\left(\sqrt{\tau}\right) \nabla f(\hat{x})$$

satisfy **the inductive hypothesis**

Algorithm. Fast Gradient Method

- **Define** $x_0 = 0$, $\tau_0 = 1$, and $z_1 = x_0 - g_0$:

$$\tau_0(f(x_0) - f_\star) + \frac{1}{2}\|z_1 - x_\star\|^2 \leq \frac{1}{2}\|x_0 - x_\star\|^2$$

- For $n = 1, \ldots, N$
  - **Apply Lemma** to $(x_{n-1}, \tau_{n-1}, z_n)$ to get (x_n, τ_n, z_{n+1}):

$$\tau_n(f(x_n) - f_\star) + \frac{1}{2}\|z_{n+1} - x_\star\|^2 \leq \frac{1}{2}\|x_0 - x_\star\|^2$$

- Output x_N

Big-O Optimality of FGM

Theorem. [Nesterov 83]

Using the FGM strategy, **Alice** can guarantee payment $\leq \frac{4}{(N+1)^2}$

Theorem. [Nemivoski Yudin 83]

If $d \geq N + 2$, no strategy for **Alice** can guarantee payment $o(\frac{1}{N^2})$

- **FGM is optimal up to constants**
- Can we do better?

Optimized Gradient Method (OGM)

- **Given**: $(x_{n-1}, \tau_{n-1}, z_n)$ satisfying

$$f_{n-1}^+ - f(x_\star) + \frac{1}{2\tau_{n-1}} \|z_n - x_\star\|^2 \leq \frac{1}{2\tau_{n-1}} \|x_\star\|^2$$

- **Idea**: Two cases
  - $f_{n-1}^+ - f(x_\star) \approx \frac{1}{2\tau_{n-1}} \|x_\star\|^2 \longrightarrow$ set $x_n = z_{n+1}$
  - $\|z_{n+1} - x_\star\|^2 \approx \|x_\star\|^2 \longrightarrow$ set $x_n = x_{n-1}^+$

  Pick $x_n = (1-\alpha)x_{n-1}^+ + \alpha z_n$ to greedily minimize

$$\min_{\alpha} \sup_{f \in \mathcal{F}} \left\{ \frac{f(x_n^+) - f(x_\star)}{\frac{1}{2}\|x_\star\|^2} : f_{n-1}^+ - f(x_\star) + \frac{1}{2\tau_{n-1}} \|z_n - x_\star\|^2 \leq \frac{1}{2\tau_{n-1}} \|x_\star\|^2 \right\}$$

- The optimal choice of $\alpha(\tau_{n-1})$ recovers OGM
- Apply "upgrade" lemma to continue induction

[Drori Teboulle 12], **[Kim Fessler 16]**

OGM induction

Lemma.

Suppose (x, τ, z) and $f \in \mathcal{F}$ satisfy **the inductive hypothesis**

$$\tau(f(x)^+ - f_\star) + \frac{1}{2}\|z - x_\star\|^2 \leq \frac{1}{2}\|x_0 - x_\star\|^2$$

Then,

$$\hat{x} = \text{some convex combination of } x^+ \text{ and } z$$
$$\hat{\tau} = \tau + \Theta\left(\sqrt{\tau}\right)$$
$$\hat{z} = z - \Theta\left(\sqrt{\tau}\right)\nabla f(\hat{x})$$

satisfy **the inductive hypothesis**

This is 99.99% true and is the right intuition.

Algorithm. Optimized Gradient Method

- **Define** $x_0 = 0$, $\tau_0 = 2$, and $z_1 = x_0 - 2g_0$:

$$\tau_0(f_0^+ - f_\star) + \frac{1}{2}\|z_1 - x_\star\|^2 \leq \frac{1}{2}\|x_0 - x_\star\|^2$$

- For $n = 1, \ldots, N-1$
  - **Apply Lemma** to $(x_{n-1}, \tau_{n-1}, z_n)$ to get (x_n, τ_n, z_{n+1}):

$$\tau_n(f_n^+ - f_\star) + \frac{1}{2}\|z_{n+1} - x_\star\|^2 \leq \frac{1}{2}\|x_0 - x_\star\|^2$$

- Slight modification for iteration N ...
- Output x_N

Minimax Optimality of OGM

Theorem. [Kim Fessler 16]

Using the OGM strategy, **Alice** can guarantee payment $\leq \frac{1}{\tau_N} \approx \frac{2}{N^2}$

Theorem. [Drori 17]

If $d \geq N + 2$, no strategy for **Alice** can guarantee payment $< \frac{1}{\tau_N}$

- **OGM is minimax optimal!**

$$\min_{\text{Alice's strategy}} \max_{\text{Bob's strategy}} (\text{Alice's payment})$$

- Can we do better?

Minimax Optimal Strategies vs. Subgame Perfect Strategies

The Pizza Line Game

- **Bob** and **Alice** are in line to grab pizza
- There are 6 slices remaining
- **Bob** goes first and may take $b \leq 3$ slices
- **Alice** goes second and may take $a \leq 6 - b$ slices
- **Alice**'s payoff $= a$ and **Bob**'s payoff $= b$
- **Consider**: **Alice** takes $a = 3$ slices
  - This strategy guarantees $a \geq 3$
  - No strategy for **Alice** can guarantee $a > 3$
- This strategy is maximin optimal

Subgame Perfect Strategies

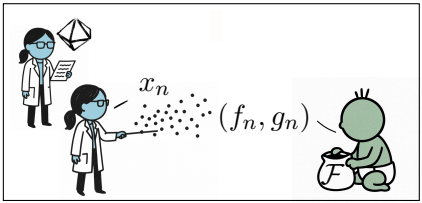
- In a sequential game, should demand a subgame perfect strategy
- A subgame is the remainder of game after some actions have been fixed
- Can identify subgame with sequence of committed actions
- In the seminar pizza line game:
  - $\{\}$
  - $\{b = 0\}, \ldots, \{b = 3\}$
- A strategy for **Alice** is **subgame perfect** if for every subgame:

$$\textbf{Alice}\text{'s subgame payoff} = \max_{\textbf{Alice}\text{'s subgame strategy}} (\textbf{Alice}\text{'s subgame payoff})$$

- The only subgame perfect strategy for **Alice** is $a = 6 - b$

Should optimizers care about subgame perfect?

- Natural notion of optimality in sequential games
- f *is not adversarial* $\iff$ Bob plays *suboptimally*
- At iteration n, **Alice** should capitalize on suboptimal play by **Bob** in iterations $0, \ldots, n-1$
- **Strong numerical advantage**
  [Luner Grimmer 25]



Subgame Perfect Gradient Method (SPGM)

[Grimmer Shu Wang 24b]

Modifying the Optimized Gradient Method

Algorithm. OGM Template

- **Initialize** $x_0 = 0$, $\tau_0 = 2$, $z_1 = x_0 - 2g_0$:

$$\tau_0(f_0^+ - f_\star) + \frac{1}{2}\|z_1 - x_\star\|^2 \leq \frac{1}{2}\|x_0 - x_\star\|^2$$

- For $n = 1, \ldots, N-1$
  - Let $(\tilde{x}, \tilde{\tau}, \tilde{z})$ be any triple satisfying the **inductive hypothesis**
  - **Apply OGM update** to $(\tilde{x}, \tilde{\tau}, \tilde{z})$ to get (x_n, τ_n, z_{n+1})

$$\tau_n(f_n^+ - f_\star) + \frac{1}{2}\|z_{n+1} - x_\star\|^2 \leq \frac{1}{2}\|x_0 - x_\star\|^2$$

- Slight modification for iteration N, Output x_N

SPGM Subproblem

- In iteration n, we have observed

$$\mathcal{H} = \{(x_0, f_0, g_0), \ldots, (x_{n-1}, f_{n-1}, g_{n-1})\}$$

and know

$$f \in \mathcal{F}_{\mathcal{H}} \coloneqq \left\{ f \in \mathcal{F} : \begin{array}{ll} f(x_i) = f_i & \forall i \in [0, n-1] \\ \nabla f(x_i) = g_i & \forall i \in [0, n-1] \end{array} \right\}$$

- **Goal**: find $(\tilde{x}, \tilde{z}, \tilde{\tau})$ to solve

$$\max_{\tilde{x}, \tilde{z} \in \mathbb{R}^d, \tilde{\tau} \in \mathbb{R}} \left\{ \tilde{\tau} : \quad \tilde{\tau}(f(\tilde{x})^+ - f_\star) + \frac{1}{2}\|\tilde{z} - x_\star\|^2 \leq \frac{1}{2}\|x_0 - x_\star\|^2 \quad \forall f \in \mathcal{F}_{\mathcal{H}} \right\}$$

- After reformulations, this is a “small” structured convex optimization problem (strictly feasible, strong duality holds)

Optimality of SPGM

Theorem.

Let $0 \le n \le N$ and consider subgame

$$\mathcal{H} = \{(x_0, f_0, g_0), \ldots, (x_{n-1}, f_{n-1}, g_{n-1})\}$$

By playing the **SPGM** strategy in this subgame, **Alice** can guarantee

$$\textbf{Alice}\text{'s subgame payment} \le \frac{1}{\tau_N(\mathcal{H})}$$

Theorem.

Suppose $d \ge N + 2$, $0 \le n \le N$, and consider subgame

$$\mathcal{H} = \{(x_0, f_0, g_0), \ldots, (x_{n-1}, f_{n-1}, g_{n-1})\}$$

No strategy for **Alice** can guarantee

$$\textbf{Alice}\text{'s subgame payment} < \frac{1}{\tau_N(\mathcal{H})}$$

- SPGM is subgame perfect!

[Grimmer Shu Wang 24]

Computational and storage overhead

- nth iter. of SPGM solves a "nonneg. least squares problem" in n variables $\longrightarrow O(nd)$ storage and $O(nd + n^{3.5})$ time
- Alternatively, impose memory constraint to $O(k)$ vectors of information. Replace

$$\tilde{\mathcal{F}}_{\mathcal{H}} \approx \left\{ f \in \mathcal{F} : \begin{array}{ll} f(x_i) = f_i & \forall i \in [n-k, n-1] \\ \nabla f(x_i) = g_i & \forall i \in [n-k, n-1] \end{array} \right\}$$

- Find $(\tilde{x}, \tilde{z}, \tilde{\tau})$ to solve

$$\max_{\tilde{x}, \tilde{z} \in \mathbb{R}^d, \tilde{\tau} \in \mathbb{R}} \left\{ \tilde{\tau} : \quad \tilde{\tau}(f(\tilde{x})^+ - f_\star) + \frac{1}{2} \|\tilde{z} - x_\star\|^2 \leq \frac{1}{2} \|x_0 - x_\star\|^2 \quad \forall f \in \tilde{\mathcal{F}}_{\mathcal{H}} \right\}$$

- $O(kd)$ storage and $O(kd + k^{3.5})$ time per iteration

Acceleration via Long Step Gradient Descent

[Grimmer Shu Wang 23, 24a]

Gradient Descent with Variable Stepsizes

- **Alice** plays a strategy of the form

$$x_{n+1} = x_n - h_n g_n$$

  for some fixed sequence $\mathbf{h} = [h_0, \ldots, h_{N-1}] \in \mathbb{R}^N$
- Standard gradient descent is $h_n = 1$ for all n
- Per-iteration descent property:

$$f_{n+1} < f_n$$

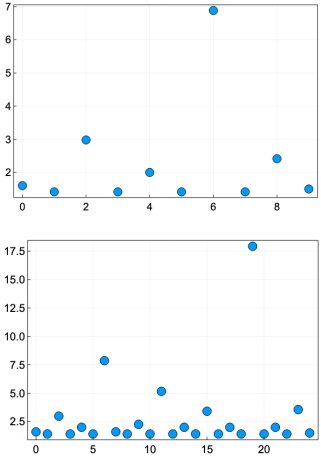
  guaranteed iff $h_n \in (0, 2)$
- Classically, the literature has focused on Short Stepsize Regime

Gradient Descent with Long Stepsizes

- **[Altschuler 18] [Daccache 19] [Eloi 22]**:
  Identification of optimal h for $N = 1, 2, 3$
- **[Das Gupta, Van Parys, Ryu '23]**:
  Numerical identification of optimal* h for $N = 1, \ldots, 25$
- **[Altschuler Parrilo 23a,b] [Grimmer Shu Wang 24a,b] [Zhang Jiang 24]**:
  Identification of conjectured optimal h for all N

$$\text{Alice's payment} = O\left(\frac{1}{N^{1.2716}}\right)$$

where $1.2716 = \log_2(1 + \sqrt{2})$



Can we accelerate by only varying stepsizes? (without momentum)

- **Given**: $(x_{n-1}, \tau_{n-1}, z_n)$ satisfying

$$f_{n-1}^+ - f(x_\star) + \frac{1}{2\tau_{n-1}} \|z_n - x_\star\|^2 \leq \frac{1}{2\tau_{n-1}} \|x_\star\|^2$$

- **Idea**: Attempt to replicate

$$x_n = (1-\alpha)x_{n-1}^+ + \alpha z_n \stackrel{?}{=} x_{n-1} - \beta g_{n-1}$$

- This works if $z_n \in x_{n-1} + \operatorname{span} g_{n-1}$

Definition.

A stepsize schedule $\mathbf{h}$ is s-composable with rate τ if

$$f_N - f_\star \leq \frac{1}{\tau}\left(\frac{1}{2}\|x_\star\|^2 - \frac{1}{2}\|z_{N+1} - x_\star\|^2\right)$$

where $z_{N+1} = x_N - \tau g_N$

Compositional properties

- $\mathbf{h} = [\,]$ is s-composable with rate 1
- Can define a composition operator $\bowtie$ so that if $\mathbf{h}_{\text{left}}$ and $\mathbf{h}_{\text{right}}$ are both s-composable with rate τ, then

$$\mathbf{h}_{\text{left}} \bowtie \mathbf{h}_{\text{right}} = [\mathbf{h}_{\text{left}}, \mu, \mathbf{h}_{\text{right}}]$$

  is s-composable with rate $(1 + \sqrt{2})\tau$

Theorem.

For any $N \geq 0$, there exists a stepsize schedule $\mathbf{h} \in \mathbb{R}^N$ so that

$$f_N - f_\star \leq O\left(\frac{1}{N^{\log_2(1+\sqrt{2})}}\right) \cdot \frac{1}{2}\|x_\star\|^2$$

- Analysis is *not* greedy!

Summary

Summary + Pointers

- Rederived as "greedy algorithms"
  - gradient descent, fast gradient method, optimized gradient descent
- Introduced subgame perfect language for formally discussing optimal dynamic guarantees: non-adversarialness of $f \in \mathcal{F}$
- Subgame perfect Gradient Method (**SPGM**): refinement of OGM that achieves subgame perfect notion
- Follow-up work:
  - Subgame perfect methods for convex **nonsmooth** optimization
    [Grimmer Wang 25]

Thank you for listening! Questions?

References I

Altschuler, J. (2018). Greed, hedging, and acceleration in convex optimization. Master's thesis, Massachusetts Institute of Technology.

Altschuler, J. M. and Parrilo, P. A. (2023a). Acceleration by stepsize hedging i: Multi-step descent and the silver stepsize schedule.

Altschuler, J. M. and Parrilo, P. A. (2023b). Acceleration by stepsize hedging ii: Silver stepsize schedule for smooth convex optimization.

Daccache, A. (2019). Performance estimation of the gradient method with fixed arbitrary step sizes. Master's thesis, Université Catholique de Louvain.

Drori, Y. and Teboulle, M. (2012). Performance of first-order methods for smooth convex minimization: a novel approach. *Mathematical Programming*, 145:451–482.

Eloi, D. (2022). Worst-case functions for the gradient method with fixed variable step sizes. Master's thesis, Université Catholique de Louvain.

Grimmer, B., Shu, K., and Wang, A. L. (2023). Accelerated gradient descent via long steps.

Grimmer, B., Shu, K., and Wang, A. L. (2024a). Accelerated objective gap and gradient norm convergence for gradient descent via long steps.

References II

Grimmer, B., Shu, K., and Wang, A. L. (2024b). Beyond minimax optimality: A subgame perfect gradient method. *arXiv preprint arXiv:2412.06731*.

Grimmer, B., Shu, K., and Wang, A. L. (2024c). Composing optimized stepsize schedules for gradient descent. *arXiv preprint arXiv:2410.16249*.

Gupta, S. D., Parys, B. P. V., and Ryu, E. (2023). Branch-and-bound performance estimation programming: A unified methodology for constructing optimal optimization methods. *Mathematical Programming*.

Kim, D. and Fessler, J. A. (2016). Optimized first-order methods for smooth convex minimization. *Math. Program.*, 159(1–2):81–107.

Nemirovskij, A. S. and Yudin, D. B. (1983). Problem complexity and method efficiency in optimization.

Nesterov, Y. (1983). A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Proceedings of the USSR Academy of Sciences*, 269:543–547.

Zhang, Z. and Jiang, R. (2024). Accelerated gradient descent by concatenation of stepsize schedules.