Algorithms and Convergence

Pontus Giselsson

Outline

- **Algorithm overview**
- Convergence and convergence rates
- Proving convergence rates

What is an algorithm?

- We are interested in algorithms that solve composite problems

$$\underset{x}{\text{minimize}}\, f(x) + g(x)$$

- An algorithm:
  - generates a sequence $(x_k)_{k\in\mathbb{N}}$ that hopefully converges to solution
  - often creates next point in sequence according to

    $$x_{k+1} = \mathcal{A}_k x_k$$

    where
    - $\mathcal{A}_k$ is a mapping that gives the next point from the current
    - $\mathcal{A}_k = \mathrm{prox}_{\gamma_k g}(I - \gamma_k \nabla f)$ for proximal gradient method

Deterministic and stochastic algorithms

- We have deterministic algorithms

$$x_{k+1} = \mathcal{A}_k x_k$$

  that given initial x_0 will give the same sequence $(x_k)_{k\in\mathbb{N}}$
- We will also see stochastic algorithms that iterate

$$x_{k+1} = \mathcal{A}_k(\xi_k) x_k$$

  where ξ_k is a random variable that also decides the mapping
  - $(x_k)_{k\in\mathbb{N}}$ is a stochastic process, i.e., collection of random variables
  - when running the algorithm, we evaluate ξ_k and get a realization
  - different realization $(x_k)_{k\in\mathbb{N}}$ every time even if started at same x_0
- Stochastic algorithms useful although problem is deterministic

Optimization algorithm overview

- Algorithms can roughly be divided into the following classes:
  - Second-order methods
  - Quasi second-order methods
  - First-order methods
  - Stochastic and coordinate-wise first-order methods
- The first three are typically deterministic and the last stochastic
- Cost of computing one iteration decreases down the list

Second-order methods

- Solves problems using second-order (Hessian) information
- Requires smooth (twice continuously differentiable) functions
- Example: Newton's method to minimize smooth function f:

$$x_{k+1} = x_k - \gamma_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$$

- Constraints can be incorporated via barrier functions:
  - Use sequence of smooth constraint barrier functions
  - Make barriers increasingly well approximate constraint set
  - For each barrier, solve smooth problem using Newton's method
  - Resulting scheme called interior point method
  - (Can be applied to directly solve primal-dual optimality condition)
- Computational backbone: solving linear systems $O(n^3)$
- Often restricted to small to medium scale problems
- We will cover Newton's method

Quasi second-order methods

- Estimates second-order information from first-order
- Solves problems using estimated second-order information
- Requires smooth (twice continuously differentiable) functions
- Quasi-Newton method for smooth f

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

  where B_k is:
    - estimate of Hessian inverse (not Hessian to avoid inverse)
    - cheaply computed from gradient information
- Computational backbone: forming B_k and matrix multiplication
- Limited memory versions exist with cheaper iterations
- Can solve large-scale smooth problems
- Will briefly look into most common method (BFGS)

First-order methods

- Solves problems using first-order (sub-gradient) information
- Computational primitives: (sub)gradients and proximal operators
- Use gradient if function differentiable, prox if nondifferentiable
- Examples for solving $\underset{x}{\text{minimize}}\, f(x) + g(x)$
  - Proximal gradient method (requires smooth f since gradient used)

    $$x_{k+1} = \text{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k))$$

  - Douglas-Rachford splitting (no smoothness requirement)

    $$z_{k+1} = \tfrac{1}{2} z_k + \tfrac{1}{2}(2\text{prox}_{\gamma g} - I)(2\text{prox}_{\gamma f} - I) z_k$$

    and $x_k = \text{prox}_{\gamma f}(z_k)$ converges to solution
- Iteration often cheaper than second-order if function split wisely
- Can solve large-scale problems
- Will look at proximal gradient method and accelerated version

Stochastic and coordinate-wise first-order methods

- Sometimes first-order methods computationally too expensive
- Stochastic gradient methods:
  - Use stochastic approximation of gradient
  - For finite sum problems, cheaply computed approximation exists
- Coordinate-wise updates:
  - Update only one (or block of) coordinates in every iteration:
    - via direct minimization
    - via proximal gradient step
  - Can update coordinates in cyclic fashion
  - Stronger convergence results if random selection of block
  - Efficient if cost of updating one coordinate is $1/n$ of full update
- Can solve huge scale problems
- Will cover randomized coordinate and stochastic methods

Outline

- Algorithm overview
- **Convergence and convergence rates**
- Proving convergence rates

Types of convergence

- Let $x^\star$ be solution to composite problem and $p^\star = f(x^\star) + g(x^\star)$
- We will see convergence of different quantities in different settings
- For deterministic algorithms that generate $(x_k)_{k\in\mathbb{N}}$, we will see
    - Sequence convergence: $x_k \to x^\star$
    - Function value convergence: $f(x_k) + g(x_k) \to p^\star$
    - If $g = 0$, gradient norm convergence: $\|\nabla f(x_k)\|_2 \to 0$
- Convergence is stronger as we go up the list
- First two common in convex setting, last in nonconvex

Convergence for stochastic algorithms

- Stochastic algorithms described by stochastic process $(x_k)_{k\in\mathbb{N}}$
- When algorithm is run, we get realization of stochastic process
- We analyze stochastic process and will see summability, e.g., of:
  - Expected distance to solution: $\sum_{k=0}^{\infty}\mathbb{E}[\|x_k - x^\star\|_2] < \infty$
  - Expected function value: $\sum_{k=0}^{\infty}\mathbb{E}[f(x_k) + g(x_k) - p^\star] < \infty$
  - If $g = 0$, expected gradient norm: $\sum_{k=0}^{\infty}\mathbb{E}[\|\nabla f(x_k)\|_2^2] < \infty$
- Sometimes arrive at weaker conclusion, when $g = 0$, that, e.g.,:
  - Expected smallest function value: $\mathbb{E}[\min_{l\in\{0,\ldots,k\}} f(x_l) - p^\star] \to 0$
  - Expected smallest gradient norm: $\mathbb{E}[\min_{l\in\{0,\ldots,k\}} \|\nabla f(x_l)\|_2] \to 0$
- Says what happens with expected value of different quantities

Algorithm realizations – Summable case

- Will conclude that sequence of expected values containing, e.g.,:

$$\mathbb{E}[\|x_k - x^\star\|_2] \quad \text{or} \quad \mathbb{E}[f(x_k) + g(x_k) - p^\star] \quad \text{or} \quad \mathbb{E}[\|\nabla f(x_k)\|_2]$$

  is summable, where all quantities are nonnegative
- What happens with the actual algorithm realizations?
- We can make conclusions by the following result: If
  - $(Z_k)_{k\in\mathbb{N}}$ is a stochastic process with $Z_k \geq 0$
  - the sequence $\{\mathbb{E}[Z_k]\}_{k\in\mathbb{N}}$ is summable: $\sum_{k=0}^{\infty} \mathbb{E}[Z_k] < \infty$

  then almost sure convergence to 0:

$$P(\lim_{k\to\infty} Z_k = 0) = 1$$

  i.e., convergence to 0 with probability 1

Algorithm realizations – Convergent case

- Will conclude that sequence of expected values containing, e.g.,:

$$\mathbb{E}[\min_{l\in\{0,\ldots,k\}} f(x_l) - p^\star] \quad \text{or} \quad \mathbb{E}[\min_{l\in\{0,\ldots,k\}} \|\nabla f(x_l)\|_2]$$

  converges to 0, where all quantities are nonnegative
- What happens with the actual algorithm realizations?
- We can make conclusions by the following result: If
  - $(Z_k)_{k\in\mathbb{N}}$ is a stochastic process with $Z_k \geq 0$
  - the expected value $\mathbb{E}[Z_k] \to 0$ as $k \to \infty$

  then convergence to 0 in probability; for all $\epsilon > 0$

$$\lim_{k\to\infty} P(Z_k > \epsilon) = 0$$

  which is weaker than almost sure convergence to 0

Convergence rates

- We have only talked about convergence, not convergence *rate*
- Rates indicate how fast (in iterations) algorithm reaches solution
- Typically divided into:
  - Sublinear rates
  - Linear rates (also called geometric rates)
  - Quadratic rates (or more generally superlinear rates)
- Sublinear rates slowest, quadratic rates fastest
- Linear rates further divided into Q-linear and R-linear
- Quadratic rates further divided into Q-quadratic and R-quadratic

Linear rates

- A Q-linear rate with factor $\rho \in [0, 1)$ can be:

$$f(x_{k+1}) + g(x_{k+1}) - p^\star \leq \rho(f(x_k) + g(x_k) - p^\star)$$
$$\mathbb{E}[\|x_{k+1} - x^\star\|_2] \leq \rho\mathbb{E}[\|x_k - x^\star\|_2]$$

- An R-linear rate with factor $\rho \in [0, 1)$ and some $C > 0$ can be:

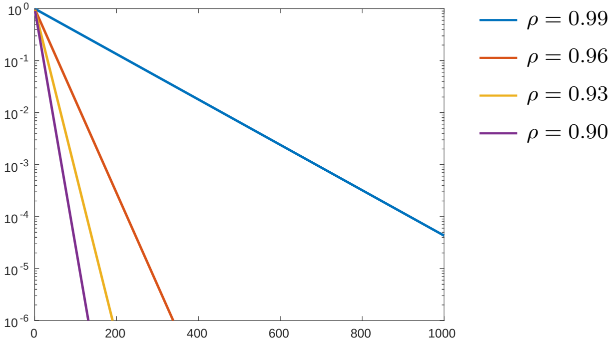
$$\|x_k - x^\star\|_2 \leq \rho^k C$$

  this is implied by Q-linear rate and has *exponential decrease*

- Linear rate is superlinear if $\rho = \rho_k$ and $\rho_k \to 0$ as $k \to \infty$
- Examples:
  - (Accelerated) proximal gradient with strongly convex cost
  - Randomized coordinate descent with strongly convex cost
  - BFGS has *local* superlinear with strongly convex cost
  - but SGD with strongly convex cost gives sublinear rate

Linear rates – Comparison

- Different rates in log-lin plot



- Called linear rate since linear in log-lin plot

Quadratic rates

- Q-quadratic rate with factor $\rho \in [0, 1)$ can be:

$$f(x_{k+1}) + g(x_{k+1}) - p^\star \leq \rho(f(x_k) + g(x_k) - p^\star)^2$$
$$\|x_{k+1} - x^\star\|_2 \leq \rho\|x - x^\star\|_2^2$$

- R-quadratic rate with factor $\rho \in [0, 1)$ and some $C > 0$ can be:

$$\|x_k - x^\star\|_2 \leq \rho^{2^k} C$$

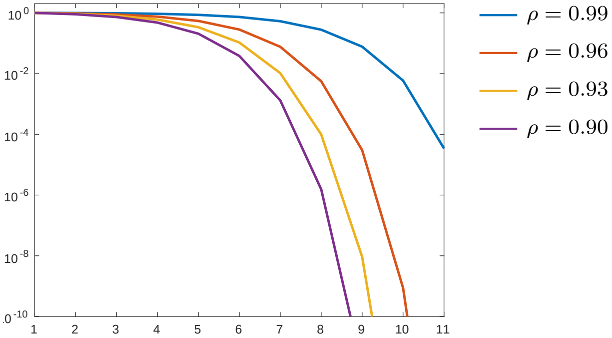
- Quadratic (ρ^{2^k}) vs linear (ρ^k) rate with factor $\rho = 0.9$:

Quadratic	Linear
1.000000000000	1.000000000000
0.810000000000	0.900000000000
0.656099945000	0.810000000000
0.430467133000	0.729000000000
0.185302002000	0.656099945000
0.034336821000	0.590490005000
0.001179017030	0.531440964000
0.000001390081	0.478296936000
0.000000000002	0.430467270000

- Example: *Locally* for Newton's method with strongly convex cost

Quadratic rates – Comparison

- Different rates in log-lin scale



- Quadratic convergence is superlinear

Sublinear rates

- A rate is sublinear if it is slower than linear
- A sublinear rate can, for instance, be of the form

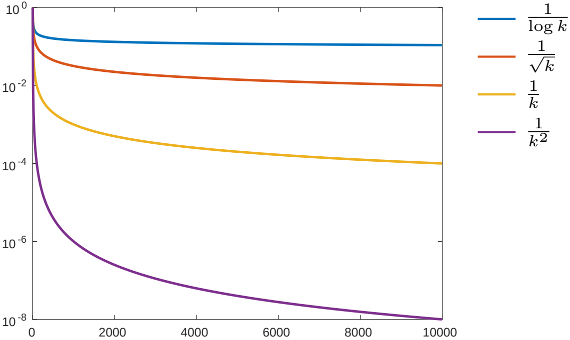
$$\begin{aligned} f(x_k) + g(x_k) - p^\star &\leq \tfrac{C}{\psi(k)} \\ \|x_{k+1} - x_k\|_2^2 &\leq \tfrac{C}{\psi(k)} \\ \min_{l=0,\ldots,k} \mathbb{E}[\|\nabla f(x_l)\|_2^2] &\leq \tfrac{C}{\psi(k)} \end{aligned}$$

  where $C > 0$ and ψ decides how fast it decreases, e.g.,
  - $\psi(k) = \log k$: Stochastic gradient descent $\gamma_k = c/k$
  - $\psi(k) = \sqrt{k}$: Stochastic gradient descent: optimal γ_k
  - $\psi(k) = k$: Proximal gradient, coordinate proximal gradient
  - $\psi(k) = k^2$: Accelerated proximal gradient method

  with improved rate further down the list
- We say that the rate is $O(\frac{1}{\psi(k)})$ for the different ψ
- To be sublinear ψ has slower than exponential growth

Sublinear rates – Comparison

- Different rates on log-lin scale



- Many iterations may be needed for high accuracy

Rate vs iteration cost

- Consider these classes of algorithms
  - Second-order methods
  - Quasi second-order methods
  - First-order methods
  - Stochastic and coordinate-wise first-order methods
- Rate deteriorates and iterations increase as we go down the list $\Downarrow$
- Iteration cost increases as we go up the list $\Uparrow$
- Performance is roughly (# iterations)$\times$(iteration cost)
- This gives a tradeoff when selecting algorithm
- Rough advise for problem size: small ($\Uparrow$) medium ($\Uparrow\Downarrow$) large ($\Downarrow$)

Outline

- Algorithm overview
- Convergence and convergence rates
- **Proving convergence rates**

Proving convergence rates

- To prove a convergence rate typically requires
  - Using inequalities that describe problem class
  - Using algorithm definition equalities (or inclusions)
  - Combine these to a form so that convergence can be concluded
- Linear and quadratic rates proofs conceptually straightforward
- Sublinear rates implicit via a *Lyapunov inequality*

Proving linear or quadratic rates

- If we suspect linear or quadratic convergence for $V_k \geq 0$:

$$V_{k+1} \leq \rho V_k^p$$

  where $\rho \in [0, 1)$ and $p = 1$ or $p = 2$ and V_k can, e.g., be

$$V_k = \|x_k - x^\star\|_2 \quad \text{or} \quad V_k = f(x_k) + g(x_k) - p^\star \quad \text{or} \quad V_k = \|\nabla f(x_k)\|_2$$

- Can prove by starting with V_{k+1} (or V_{k+1}^2) and continue using
  - function class inequalities
  - algorithm equalities
  - propeties of norms
  - $\ldots$

Sublinear convergence – Lyapunov inequality

- Assume we want to show sublinear convergence of some $R_k \geq 0$
- This typically requires finding a *Lyapunov inequality*:

$$V_{k+1} \leq V_k + W_k - R_k$$

  where
  - $(V_k)_{k\in\mathbb{N}}$, $(W_k)_{k\in\mathbb{N}}$, and $(R_k)_{k\in\mathbb{N}}$ are nonnegative real numbers
  - $(W_k)_{k\in\mathbb{N}}$ is summable, i.e., $\overline{W} := \sum_{k=0}^{\infty} W_k < \infty$
- Such a Lyapunov inequality can be found by using
  - function class inequalities
  - algorithm equalities
  - propeties of norms
  - ...

Lyapunov inequality consequences

- From the Lyapunov inequality:

$$V_{k+1} \leq V_k + W_k - R_k$$

  we can conclude that
  - V_k is nonincreasing if all $W_k = 0$
  - V_k converges as $k \to \infty$ (will not prove)
- Recursively applying the inequality for $l \in \{k, \ldots, 0\}$ gives

$$V_{k+1} \leq V_0 + \sum_{l=0}^{k} W_l - \sum_{l=0}^{k} R_l \leq V_0 + \overline{W} - \sum_{l=0}^{k} R_l$$

  where $\overline{W}$ is infinite sum of W_k, this implies

$$\sum_{l=0}^{k} R_l \leq V_0 - V_{k+1} + \sum_{l=0}^{k} W_l \leq V_0 + \sum_{l=0}^{k} W_l \leq V_0 + \overline{W}$$

  from which we can
  - conclude that $R_k \to 0$ as $k \to \infty$ since $R_k \geq 0$
  - derive sublinear rates of convergence for R_k towards 0

Concluding sublinear convergence

- Lyapunov inequality consequence restated

$$\sum_{l=0}^{k} R_l \leq V_0 + \sum_{l=0}^{k} W_l \leq V_0 + \overline{W}$$

- We can derive sublinear convergence for
  - Best R_k: $(k+1)\min_{l\in\{0,\ldots,k\}} R_l \leq \sum_{l=0}^{k} R_l$
  - Last R_k (if R_k decreasing): $(k+1)R_k \leq \sum_{l=0}^{k} R_l$
  - Average R_k: $\bar{R}_k = \frac{1}{k+1}\sum_{l=0}^{k} R_l$
- Let $\hat{R}_k$ be any of these quantities, and we have

$$\hat{R}_k \leq \frac{\sum_{l=0}^{k} R_l}{k+1} \leq \frac{V_0 + \overline{W}}{k+1}$$

which shows a $O(1/k)$ sublinear convergence

Deriving other than $O(1/k)$ convergence (1/3)

- Other rates can be derived from a modified Lyapunov inequality:

$$V_{k+1} \leq V_k + W_k - \lambda_k R_k$$

with $\lambda_k > 0$ when we are interested in convergence of R_k, then

$$\sum_{l=0}^{k} \lambda_l R_l \leq V_0 + \sum_{l=0}^{k} W_l \leq V_0 + \overline{W}$$

- We have $R_k \to 0$ as $k \to \infty$ if, e.g., $\sum_{l=0}^{\infty} \lambda_l = \infty$

Deriving other than $O(1/k)$ convergence (2/3)

- Restating the consequence: $\sum_{l=0}^{k} \lambda_l R_l \leq V_0 + \overline{W}$
- We can derive sublinear convergence for
  - Best R_k: $\min_{l \in \{0,\ldots,k\}} R_l \sum_{l=0}^{k} \lambda_l \leq \sum_{l=0}^{k} \lambda_l R_l$
  - Last R_k (if R_k decreasing): $R_k \sum_{l=0}^{k} \lambda_l \leq \sum_{l=0}^{k} \lambda_l R_l$
  - Weighted average R_k: $\bar{R}_k = \frac{1}{\sum_{l=0}^{k} \lambda_l} \sum_{l=0}^{k} \lambda_l R_l$
- Let $\hat{R}_k$ be any of these quantities, and we have

$$\hat{R}_k \leq \frac{\sum_{l=0}^{k} R_l}{\sum_{l=0}^{k} \lambda_l} \leq \frac{V_0 + \overline{W}}{\sum_{l=0}^{k} \lambda_l}$$

Deriving other than $O(1/k)$ convergence (3/3)

- How to get a rate out of:

$$\hat{R}_k \leq \frac{V_0 + \overline{W}}{\sum_{l=0}^{k} \lambda_l}$$

- Assume $\psi(k) \leq \sum_{l=0}^{k} \lambda_l$, then $\psi(k)$ decides rate:

$$\hat{R}_k \leq \frac{\sum_{l=0}^{k} R_l}{\sum_{l=0}^{k} \lambda_l} \leq \frac{V_0 + \overline{W}}{\psi(k)}$$

  which gives a $O(\frac{1}{\psi(k)})$ rate

  - If $\lambda_k = c$ is constant: $\psi(k) = c(k+1)$ and we have $O(1/k)$ rate
  - If λ_k is decreasing: slower rate than $O(1/k)$
  - If λ_k is increasing: faster rate than $O(1/k)$

Estimating ψ via integrals

- Assume that $\lambda_k = \phi(k)$, then $\psi(k) \le \sum_{l=0}^{k} \phi(l)$ and

$$\hat{R}_k \le \frac{\sum_{l=0}^{k} R_l}{\sum_{l=0}^{k} \phi(l)} \le \frac{V_0 + \overline{W}}{\psi(k)}$$

- To estimate ψ, we use the integral inequalities
  - for decreasing nonnegative ϕ:

$$\int_{t=0}^{k} \phi(t)dt + \phi(k) \le \sum_{l=0}^{k} \phi(l) \le \int_{t=0}^{k} \phi(t)dt + \phi(0)$$

  - for increasing nonnegative ϕ:

$$\int_{t=0}^{k} \phi(t)dt + \phi(0) \le \sum_{l=0}^{k} \phi(l) \le \int_{t=0}^{k} \phi(t)dt + \phi(k)$$

- Remove $\phi(k), \phi(0) \ge 0$ from the lower bounds and use estimate:

$$\psi(k) = \int_{t=0}^{k} \phi(t)dt \le \sum_{l=0}^{k} \phi(l)$$

Sublinear rate examples

- For Lyapunov inequality $V_{k+1} \leq V_k + W_k - \lambda_k R_k$, we get:

$$\hat{R}_k \leq \frac{V_0 + \overline{W}}{\psi(k)} \qquad \text{where} \qquad \lambda_k = \phi(k) \text{ and } \psi(k) = \int_{t=0}^{k} \phi(t)dt$$

- Let us quantify the rate ψ in a few examples:
  - Two examples that are slower than $O(1/k)$:
    - $\lambda_k = \phi(k) = c/(k+1)$ gives slow $O(\frac{1}{\log k})$ rate:

$$\psi(k) = \int_{t=0}^{k} \frac{c}{t+1} dt = c[\log(t+1)]_{t=0}^{k} = c\log(k+1)$$

    - $\lambda_k = \phi(k) = c/(k+1)^\alpha$ for $\alpha \in (0,1)$, gives faster $O(\frac{1}{k^{1-\alpha}})$ rate:

$$\psi(k) = \int_{t=0}^{k} \frac{c}{(t+1)^\alpha} dt = c[\tfrac{(t+1)^{1-\alpha}}{(1-\alpha)}]_{t=0}^{k} = \tfrac{c}{1-\alpha}((k+1)^{1-\alpha} - 1)$$

  - An example that is faster than $O(1/k)$
    - $\lambda_k = \phi(k) = c(k+1)$ gives $O(\frac{1}{k^2})$ rate:

$$\psi(k) = \int_{t=0}^{k} c(t+1)dt = c[\tfrac{1}{2}(t+1)^2]_{t=0}^{k} = \tfrac{c}{2}((k+1)^2 - 1)$$

Stochastic setting and law of total expectation

- In the stochastic setting, we analyze the stochastic process

$$x_{k+1} = \mathcal{A}_k(\xi_k)x_k$$

- We will look for inequalities of the form

$$\mathbb{E}[V_{k+1}|x_k] \leq \mathbb{E}[V_k|x_k] + \mathbb{E}[W_k|x_k] - \lambda_k \mathbb{E}[R_k|x_k]$$

  to see what happens in one step given x_k (but not given ξ_k)
- We use *law of total expectation* $\mathbb{E}[\mathbb{E}[X|Y]] = \mathbb{E}[X]$ to get

$$\mathbb{E}[V_{k+1}] \leq \mathbb{E}[V_k] + \mathbb{E}[W_k] - \lambda_k \mathbb{E}[R_k]$$

  which is a Lyapunov inequality
- We can draw rate conclusions, as we did before, now for $\mathbb{E}[R_k]$
- For realizations we can say:
  - If $\mathbb{E}[R_k]$ is summable, then $R_k \to 0$ almost surely
  - If $\mathbb{E}[R_k] \to 0$, then $R_k \to 0$ in probability

Rates in stochastic setting

- Lyapunov inequality $\mathbb{E}[V_{k+1}] \leq \mathbb{E}[V_k] + \mathbb{E}[W_k] - \lambda_k \mathbb{E}[R_k]$ implies:

$$\sum_{l=0}^{k} \lambda_l \mathbb{E}[R_l] \leq V_0 + \sum_{l=0}^{\infty} \mathbb{E}[W_l] \leq V_0 + \bar{W}$$

- Same procedure as before gives sublinear rates for
  - Best $\mathbb{E}[R_k]$: $\min_{l\in\{0,\ldots,k\}} \mathbb{E}[R_l] \sum_{l=0}^{k} \lambda_l \leq \sum_{l=0}^{k} \lambda_l \mathbb{E}[R_l]$
  - Last $\mathbb{E}[R_k]$ (if $\mathbb{E}[R_k]$ decreasing): $\mathbb{E}[R_k] \sum_{l=0}^{k} \lambda_l \leq \sum_{l=0}^{k} \lambda_l \mathbb{E}[R_l]$
  - Weighted average: $\mathbb{E}[\bar{R}_k] = \frac{1}{\sum_{l=0}^{k} \lambda_l} \sum_{l=0}^{k} \lambda_l \mathbb{E}[R_l]$
- Jensen's inequality for concave $\min_l$ in best residual reads

$$\mathbb{E}[\min_{l\in\{0,\ldots,k\}} R_l] \leq \min_{l\in\{0,\ldots,k\}} \mathbb{E}[R_l]$$

- Let $\hat{R}_k$ be any of the above quantities, and we have

$$\mathbb{E}[\hat{R}_k] \leq \frac{V_0 + \bar{W}}{\sum_{l=0}^{k} \lambda_l}$$