Optimization in Machine Learning

Lecture 14: Algorithms for Optimization, Convergence Analysis of Gradient Descent under Lipschitz Continuity and Convexity, Enhancements via Smoothness and Strong Convexity

Ganesh Ramakrishnan

Department of Computer Science
Dept of CSE, IIT Bombay
https://www.cse.iitb.ac.in/~ganesh

February, 2025

[Recap] Convergence rate for Convexity + Lipschitz Continuity

- Define $\hat{x} = \text{argmin}_i f(x_i)$. Then,

$$|f(\hat{x}) - f(x^*)| \leq \frac{RB}{\sqrt{T}}$$

- If we need $|f(\hat{x}) - f(x^*)| \leq \epsilon$, it suffices to have

$$\frac{RB}{\sqrt{T}} \leq \epsilon$$

- Which implies that:

$$T \geq \frac{R^2 B^2}{\epsilon^2}$$

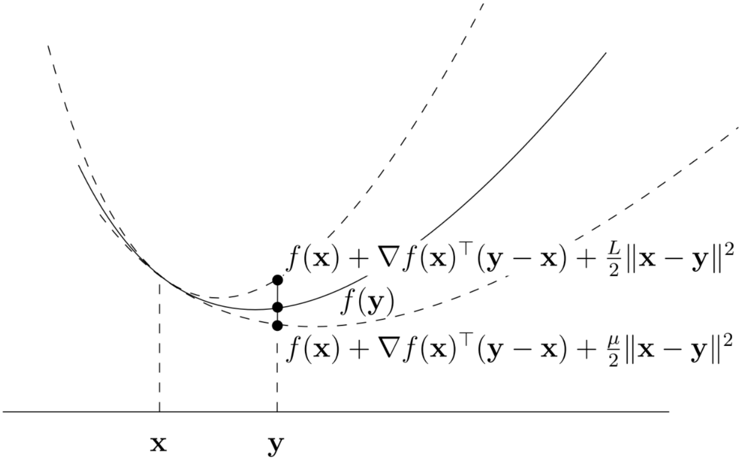
- Final Result: Given a Lipschitz continuous function f, gradient descent with step size $\gamma = \frac{R}{B\sqrt{T}}$ achieves a solution $\hat{x}$ s.t $|f(\hat{x}) - f(x^*)| \leq \epsilon$ in $\frac{R^2 B^2}{\epsilon^2}$ iterations.

[Recap] Convergence rate for Convexity + Lipschitz Smoothness

- Putting everything together: $f(x_T) - f(x^*) \leq \frac{L}{2T}||x_0 - x^*||^2 = \frac{LR^2}{2T}$
- To ensure that $f(x_T) - f(x^*) \leq \epsilon$, we require $\frac{LR^2}{2T} \leq \epsilon$.
- This implies that $T \geq \frac{R^2 L}{2\epsilon}$
- To achieve an error of 0.01, we require $50R^2L$ iterations instead of $10^4R^2B^2$ in the Lipschitz case!
- Final Result: Given a L smooth convex function f, Gradient descent with step size $\gamma = \frac{1}{L}$ achieves a solution x_T s.t $|f(x_T) - f(x^*)| \leq \epsilon$ in $\frac{R^2L}{\epsilon}$ iterations.

Recall this value was to give a lowest upper bound
In practice line/ray search techniques are used and
convergence can be proved with Strong Wolfe conditions on step size

Smooth + Strongly Convex Functions



Source: Martin Jaggi (CS 439)

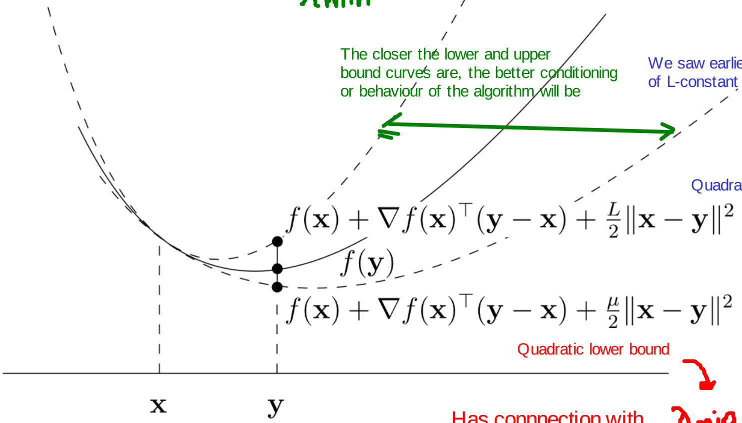
Smooth + Strongly Convex Functions

$\frac{\lambda_{max}}{\lambda_{min}}$: Associated with the condition number of the (Hessian) matrix (of say Logistic Loss)

The closer the lower and upper bound curves are, the better conditioning or behaviour of the algorithm will be

We saw earlier today the connection of L-constant for L-continuity with λ_{max}

Quadratic Upper bound



$(\nabla^2 f \succeq \mu I)$

Quadratic lower bound

Has connnection with λ_{min}

$\mu \le \lambda_{min}$ & $\lambda_{max} \le L$ ~ $\frac{\lambda_{max}}{\lambda_{min}} \le \frac{L}{\mu}$

Source: Martin Jaggi (CS 439)

Fastest Convergence with Strong Convexity (+ Smoothness later) I

- Recall from Analysis I, (1) based on straightforward algebra:

$$g_t^T(x_t - x^*) = \frac{\gamma_t}{2}||g_t||^2 + \frac{1}{2\gamma_t}\left(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2\right) \tag{1}$$

TRICK 1 Used

Deviation: Instead of convexity followed by telescopic summing, why not STRONG convexity next...

Homework: How do we use strong convexity in conjunction with L-smothness to get a sufficient condition as

T >= log(1/ϵ)

Can it be through some intermediate steps culminating in

f(μ/L)^T <= ϵ

?

Fastest Convergence with Strong Convexity (+ Smoothness later) I

- Recall from Analysis I, (1) based on straightforward algebra:

$$g_t^T(x_t - x^*) = \frac{\gamma_t}{2}||g_t||^2 + \frac{1}{2\gamma_t}\left(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2\right) \tag{1}$$

- We can use a stronger lower bound on the above expression(s) via strong convexity:

$$g_t^T(x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\mu}{2}||x_t - x^*||^2 \tag{2}$$

Fastest Convergence with Strong Convexity (+ Smoothness later) I

- Recall from Analysis I, (1) based on straightforward algebra:

$$g_t^T(x_t - x^*) = \frac{\gamma_t}{2}||g_t||^2 + \frac{1}{2\gamma_t}\left(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2\right) \tag{1}$$

- We can use a stronger lower bound on the above expression(s) via strong convexity:

$$g_t^T(x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\mu}{2}||x_t - x^*||^2 \tag{2}$$

- Putting together (1) and (2) and rearranging terms:

$$f(x_t) - f(x^*) \leq \frac{1}{2\gamma}(\gamma^2||g_t||^2 + ||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) - \frac{\mu}{2}||x_t - x^*||^2$$

$$\Rightarrow ||x_{t+1} - x^*||^2 \leq 2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$$

Fastest Convergence with Strong Convexity (+ Smoothness later) II

- From previous slide:
  $||x_{t+1} - x^*||^2 \leq \boxed{2\gamma\left(f(x^*) - f(x_t)\right)} + \boxed{\gamma^2||g_t||^2} + \boxed{(1 - \mu\gamma)||x_t - x^*||^2}$

Fastest Convergence with Strong Convexity (+ Smoothness later) II

- From previous slide:

$$||x_{t+1} - x^*||^2 \leq 2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$$

Telescopic summing after dividing both sides by $(1-\mu r)^{t+1}$?

Will it help if we somehow show that this is ≤ 0

$$\sum_{t=0}^{T-1} \frac{||x_{t+1}-x^*||^2}{(1-\mu r)^{t+1}} \leq \sum_{t=0}^{T-1} \frac{2r(f(x^*)-f(x_t)) + r^2||g_t||^2}{(1-\mu r)^{t+1}} + \frac{||x_t-x^*||^2}{(1-\mu r)^t}$$

$$\frac{||x_T-x^*||^2}{(1-\mu r)^T} \leq \sum_{t=0}^{T-1} \frac{2r(f(x^*)-f(x_t)) + r^2||g_t||^2}{(1-\mu r)^{t+1}} + ||x_0-x^*||^2 \leq ||x_0-x^*||$$

Recall guaranteed
decrease based on L-smoothness
Next few slides we prove this is <=0

Fastest Convergence with Strong Convexity (+ Smoothness later) II

- From previous slide:

$$\|x_{t+1} - x^*\|^2 \leq 2\gamma(f(x^*) - f(x_t)) + \gamma^2\|g_t\|^2 + (1 - \mu\gamma)\|x_t - x^*\|^2$$

- Recall (from previous Lecture), for Lipschitz smooth function f,

$$\boxed{f(x_{t+1}) \leq f(x_t) + g_t^T(x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2 \quad \leq f(x_t) - \gamma\|g_t\|^2 + \frac{L}{2}\gamma^2\|g_t\|^2 \quad (3)}$$

Will it help us simplify this expression?

$$\sum_{t=0}^{T-1} \frac{2\gamma(f(x^*) - f(x_t)) + \gamma^2\|g_t\|^2}{(1-\mu\gamma)^{t+1}}$$

Note: $f(x^*) \leq f(x_t)$

Should we substitute $\gamma = 1/L$?

Recall: We had also minimized RHS wrt γ (for $\gamma = \frac{1}{L}$)
$f(x_t) - \frac{1}{2L}\|g_t\|^2$

Fastest Convergence with Strong Convexity (+ Smoothness later) II

- From previous slide:
  $||x_{t+1} - x^*||^2 \leq 2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$
- Recall (from previous Lecture), for Lipschitz smooth function f,

$$f(x_{t+1}) \leq f(x_t) + g_t^T(x_{t+1} - x_t) + \frac{L}{2}||x_{t+1} - x_t||^2 \quad \leq f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2 \quad (3)$$

- For $\gamma = 1/L$, $f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2$ is minimized and since $\boxed{f(x^*) \leq f(x_{t+1}),}$

$$f(x^*) \leq f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||g_t||^2$$

Fastest Convergence with Strong Convexity (+ Smoothness later) II

- From previous slide:
  $||x_{t+1} - x^*||^2 \leq 2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 + (1-\mu\gamma)||x_t - x^*||^2$
- Recall (from previous Lecture), for Lipschitz smooth function f,

$$f(x_{t+1}) \leq f(x_t) + g_t^T(x_{t+1} - x_t) + \frac{L}{2}||x_{t+1} - x_t||^2 \quad \leq f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2 \quad (3)$$

- For $\gamma = 1/L$, $f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2$ is minimized and since $f(x^*) \leq f(x_{t+1})$,

$$f(x^*) \leq f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||g_t||^2$$

- Now let us show that $2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 \leq 0$. For step size $\gamma = 1/L$.

$$2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 = \frac{2}{L}(f(x^*) - f(x_t)) + \frac{1}{L^2}||g_t||^2$$

$$\leq -\frac{1}{L^2}||g_t||^2 + \frac{1}{L^2}||g_t||^2$$

Fastest Convergence with Strong Convexity (+ Smoothness later) II

- From previous slide:

$$||x_{t+1} - x^*||^2 \le 2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 + (1-\mu\gamma)||x_t - x^*||^2$$

- Recall (from previous Lecture), for Lipschitz smooth function f,

$$f(x_{t+1}) \le f(x_t) + g_t^T(x_{t+1} - x_t) + \frac{L}{2}||x_{t+1} - x_t||^2 \le f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2 \quad (3)$$

- For $\gamma = 1/L$, $f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2$ is minimized and since $f(x^*) \le f(x_{t+1})$,

$$\boxed{f(x^*) \le f(x_{t+1}) \le f(x_t) - \frac{1}{2L}||g_t||^2}$$

- Now let us show that $2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 \le 0$. For step size $\boxed{\gamma = 1/L.}$

$$2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 = \boxed{\frac{2}{L}\left(f(x^*) - f(x_t)\right)} + \frac{1}{L^2}||g_t||^2$$

$$\le \boxed{-\frac{1}{L^2}||g_t||^2} + \frac{1}{L^2}||g_t||^2$$

Fastest Convergence with Smooth + Strongly Convex (now) II

- Since: $||x_{t+1} - x^*||^2 \leq 2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$

Fastest Convergence with Smooth + Strongly Convex (now) II

- Since: $||x_{t+1} - x^*||^2 \leq 2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 + (1-\mu\gamma)||x_t - x^*||^2$
- And for Lipschitz smooth function f,

$$2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 = \frac{2}{L}\left(f(x^*) - f(x_t)\right) + \frac{1}{L^2}||g_t||^2$$

$$\leq -\frac{1}{L^2}||g_t||^2 + \frac{1}{L^2}||g_t||^2$$

Fastest Convergence with Smooth + Strongly Convex (now) II

- Since: $||x_{t+1} - x^*||^2 \leq 2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 + (1-\mu\gamma)||x_t - x^*||^2$
- And for Lipschitz smooth function f,

$$2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 = \frac{2}{L}\left(f(x^*) - f(x_t)\right) + \frac{1}{L^2}||g_t||^2$$

$$\leq -\frac{1}{L^2}||g_t||^2 + \frac{1}{L^2}||g_t||^2$$

- We have, by packing everything together:

$$||x_{t+1} - x^*||^2 \leq \left(1 - \frac{\mu}{L}\right)||x_t - x^*||^2$$

Fastest Convergence with Smooth + Strongly Convex (now) II

- Since: $||x_{t+1} - x^*||^2 \leq 2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 + (1-\mu\gamma)||x_t - x^*||^2$
- And for Lipschitz smooth function f,

$$2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 = \frac{2}{L}\left(f(x^*) - f(x_t)\right) + \frac{1}{L^2}||g_t||^2$$
$$\leq -\frac{1}{L^2}||g_t||^2 + \frac{1}{L^2}||g_t||^2$$

- We have, by packing everything together:

$$||x_{t+1} - x^*||^2 \leq \underbrace{\left(1 - \frac{\mu}{L}\right)}_{\in (0,1)}||x_t - x^*||^2$$

Recall $\mu I \leq \nabla^2 f \leq LI$

$\Rightarrow \mu \leq L$

Fastest Convergence with Smooth + Strongly Convex III

- Packing everything together:

$$||x_{t+1} - x^*||^2 \leq \boxed{\left(1 - \frac{\mu}{L}\right) ||x_t - x^*||^2}$$

Fastest Convergence with Smooth + Strongly Convex III

- Packing everything together:

$$||x_{t+1} - x^*||^2 \leq \left(1 - \frac{\mu}{L}\right) ||x_t - x^*||^2$$

Finally this is Q-linearly convergent in x (not in f yet)

- $v^1, \ldots, v^k$ is **Q-linearly** convergent if

$$\frac{\|v^{k+1} - v^*\|}{\|v^k - v^*\|} \leq r$$

for some $k \geq \theta$, and $r \in (0, 1)$

Multiply & have what we jumped to earlier

$$\frac{||x_T - x^*||^2}{(1-\mu/L)^T} \leq ||x_0 - x^*||^2$$

Fastest Convergence with Smooth + Strongly Convex III

- Packing everything together:

$$||x_{t+1} - x^*||^2 \leq \boxed{\left(1 - \frac{\mu}{L}\right) ||x_t - x^*||^2}$$

- Multiplying all terms from $t = 0, \cdots, T-1$:

$$||x_T - x^*||^2 \leq \boxed{\left(1 - \frac{\mu}{L}\right)^T} ||x_0 - x^*||^2$$

Fastest Convergence with Smooth + Strongly Convex III

Earlier: $T = O\left(\frac{1}{\epsilon}\right)$

- Packing everything together:

$$||x_{t+1} - x^*||^2 \leq \boxed{\left(1 - \frac{\mu}{L}\right)} ||x_t - x^*||^2$$

- Multiplying all terms from $t = 0, \cdots, T-1$:

$$||x_T - x^*||^2 \leq \boxed{\left(1 - \frac{\mu}{L}\right)^T} ||x_0 - x^*||^2$$

Hint to $T = O\left(\log \frac{1}{\epsilon}\right)$

What we seek is $f(x_T) \to f(x^*)$

Fastest Convergence with Smooth + Strongly Convex III

- Packing everything together:

$$||x_{t+1} - x^*||^2 \leq \left(1 - \frac{\mu}{L}\right) ||x_t - x^*||^2$$

- Multiplying all terms from $t = 0, \cdots, T-1$:

$$||x_T - x^*||^2 \leq \left(1 - \frac{\mu}{L}\right)^T ||x_0 - x^*||^2$$

- In our final step, we combine smoothness and the fact that $\nabla f(x^*) = 0$:

$$f(x_T) - f(x^*) \leq \nabla f(x^*)^T (x_T - x^*) + \frac{L}{2}||x_T - x^*||^2 = \frac{L}{2}||x_T - x^*||^2$$

$$\Rightarrow f(x_T) - f(x^*) \leq \frac{L}{2}||x_T - x^*||^2 \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T ||x_0 - x^*||^2$$

Fastest Convergence with Smooth + Strongly Convex III

- Packing everything together:

$$||x_{t+1} - x^*||^2 \leq \left(1 - \frac{\mu}{L}\right) ||x_t - x^*||^2$$

- Multiplying all terms from $t = 0, \cdots, T-1$:

$$||x_T - x^*||^2 \leq \left(1 - \frac{\mu}{L}\right)^T ||x_0 - x^*||^2$$

Necessary

- In our final step, we combine smoothness and the fact that $\boxed{\nabla f(x^*) = 0}$:

$$f(x_T) - f(x^*) \leq \nabla f(x^*)^T (x_T - x^*) + \frac{L}{2}||x_T - x^*||^2 = \frac{L}{2}||x_T - x^*||^2$$

$$\Rightarrow f(x_T) - f(x^*) \leq \frac{L}{2}||x_T - x^*||^2 \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T ||x_0 - x^*||^2$$

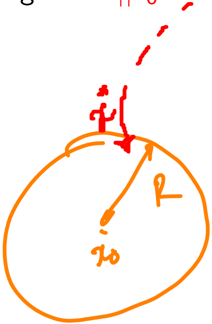
Convergence Rate For Smooth + Strongly Convex

- Setting $R^2 = ||x_0 - x^*||^2$, we get:

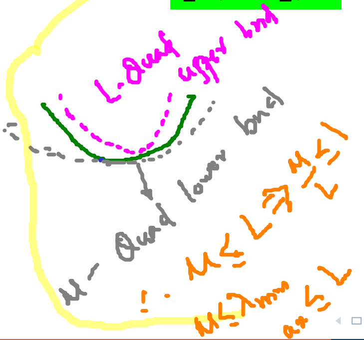
$$\boxed{f(x_T) - f(x^*)} \leq \boxed{\frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T} R^2$$

Convergence Rate For Smooth + Strongly Convex

- Setting $R^2 = ||x_0 - x^*||^2$, we get:



$$\boxed{f(x_T) - f(x^*)} \le \boxed{\frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T} R^2$$



Convergence Rate For Smooth + Strongly Convex

- Setting $R^2 = ||x_0 - x^*||^2$, we get:

$$\boxed{f(x_T) - f(x^*)} \le \boxed{\frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T} R^2$$

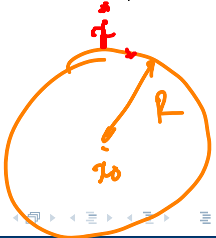
- To get an error of ϵ, we require $\boxed{\frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T} R^2 \le \epsilon$ which implies $T \ge \frac{L}{\mu}\log(\frac{R^2 L}{2\epsilon})$.

Convergence Rate For Smooth + Strongly Convex

- Setting $R^2 = ||x_0 - x^*||^2$, we get:

$$\boxed{f(x_T) - f(x^*)} \leq \boxed{\frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T} R^2$$

- To get an error of ϵ, we require $\boxed{\frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T} R^2 \leq \epsilon$ which implies $T \geq \frac{L}{\mu}\log(\frac{R^2 L}{2\epsilon})$.



Convergence Rate For Smooth + Strongly Convex

- Setting $R^2 = ||x_0 - x^*||^2$, we get:

$$f(x_T) - f(x^*) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T R^2$$

- To get an error of ϵ, we require $\frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T R^2 \leq \epsilon$ which implies $T \geq \frac{L}{\mu}\log(\frac{R^2 L}{2\epsilon})$.
- To get an error of $\epsilon = 0.01$, we now need only $L/\mu \log(50R^2L)$ iterations as opposed to $50R^2L$ iterations in the smooth case!

Summary of Results so Far...(with convexity by default)

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then, we have the following:
  - $\epsilon = 0.1$, C: 10000, S = 50, SS = 8.49 iterations
  - $\epsilon = 0.01$, C: 1000000, S = 500, SS = 13.49 iterations
  - $\epsilon = 0.001$, C: 100000000, S = 5000, SS = 18.49 iterations
- As ϵ reduces by 10, the number of iterations of strongly + smooth case increases only by a additive constant! This is linear convergence!

Summary of Results so Far...(with convexity by default)

worst case for some intermediat ineq.

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then, we have the following:
  - $\epsilon = 0.1$, C: 10000, S = 50, SS = 8.49 iterations
  - $\epsilon = 0.01$, C: $\overset{*100}{1000000}$, S = $\overset{*10}{500}$, SS = 13.49 iterations
  - $\epsilon = 0.001$, C: $\overset{*100}{100000000}$, S = $\overset{*10}{5000}$, SS = 18.49 iterations

constant step increase for 10 fold higher precision

constant step increase for 10 fold higher precision

- As ϵ reduces by 10, the number of iterations of strongly + smooth case increases only by a additive constant! This is linear convergence!

Revisist optional slides
and notice that this
is Q-linear convergence