Optimization in Machine Learning

Lecture 13: 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

- Recall final result: E_1 (which is independent of γ)

$$\boxed{\sum_{t=0}^{T-1}(f(x_t)-f(x^*))} \le \frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2+\frac{1}{2\gamma}(||x_0-x^*||^2)$$

- Let $||x_0-x^*|| \le R$ and $||\nabla f(x)|| \le B$ for all x. Then...
- $\frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2+\frac{1}{2\gamma}(||x_0-x^*||^2) \le \frac{\gamma}{2}TB^2+\frac{R^2}{2\gamma} \rightarrow E_2(\gamma)$
- The extreme right expression $\frac{\gamma}{2}TB^2+\frac{R^2}{2\gamma}$ is minimized with respect to γ, by setting its derivative (wrt γ) to 0 which is obtained by setting $\gamma=\frac{R}{B\sqrt{T}}$.

to value $\frac{RB}{\sqrt{T}}$.

Trick 3: Determine the minimum value of an upper bound (likewise maximum value of lower bound)

E_1 (which is independent of γ) $\dashrightarrow E_1 \le E_2(\gamma) \Leftrightarrow E_1 \le \min_{\gamma} E_2(\gamma)$

Note: This inequality holds for any $\gamma > 0$
However, choice of step size can be important for analysis of convergence of some optimization algorithms

- Recall final result:

E_1 (which is independent of γ) $E_2(\gamma)$

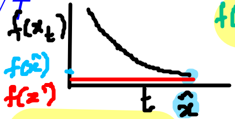
$$\boxed{\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \le \frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2)}$$

- Let $||x_0 - x^*|| \le R$ and $||\nabla f(x)|| \le B$ for all x. Setting $\gamma = \frac{R}{B\sqrt{T}}$,
- We obtain:

E_1 (which is independent of γ) $\min_{\gamma} E_2(\gamma)$

$$\frac{1}{T}[f(\hat{x}) - f(x^*)] = \frac{1}{T}\sum_{t=1}^{T-1}[f(\hat{x}) - f(x^*)] \le \frac{1}{T}\sum_{t=0}^{T-1}[f(x_t) - f(x^*)] \le \frac{RB}{\sqrt{T}}$$

1) Recap from Machine Learning that often the iterations are stopped based on validation set (i.e. when validation accuracy starts dropping) Here, the optimization is wrt training set. So we have early stopping for avoiding overfitting/to generalize well
2) Numerical precision can sometimes be a deterrent to get exact equality of $f(\hat{x}) = f(x^*)$



- Last iterate not necessarily the best!
- Choose $\hat{x} = \text{argmin}_i f(x_i)$ as the final iterate. Show that $|f(\hat{x}) - f(x^*)|$ satisfies the above bound.

[Recap] Lipschitz Continuous Functions: Final Bound

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

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

[Recap] Lipschitz Continuous Functions: Final Bound

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

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

Suppose our need is to find T such that

$$|f(\hat{x}) - \underbrace{f(x^*)}| < \epsilon$$

Note that we do not have access to this value!

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

Q: Can this help derive a sufficient condition on T?

Ans: Yes. Just flip numberator and denominator and square both $\Rightarrow \frac{R^2 B^2}{\epsilon^2} \leq T$

[Recap] Lipschitz Continuous Functions: Final Bound

- 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

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

[Recap] Lipschitz Continuous Functions: Final Bound

- 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:

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

[Recap] Lipschitz Continuous Functions: Final Bound

- 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:

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

Optional: Rate&Order of Convergence, Generalized Gradient Descent:

https://moodle.iitb.ac.in/pluginfile.php/143881/mod_resource/content/2/Optional%20Reading%20-%20Q-Convergence%20principles%2Bgeneral%20descent%20algorithms%2Bbacktracking%20ray%20search%2Barmijo%20conditions.pdf

HOMEWORK

$T = O\left(\frac{1}{\epsilon^2}\right)$?

OR

$T = \Omega\left(\frac{1}{\epsilon^2}\right)$?

This kind of analysis is less relevant for us

See R&Q-convergence

BIG Omega is about the limiting (lower bound) behaviour of the function for epsilon tending to infinity

[Recap] Lipschitz Continuous Functions: Final Bound

- 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.

How good or bad is this bound?

- Final Result: Given a B-Lipschitz continuous function convex 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.

How good or bad is this bound?

- Final Result: Given a B-Lipschitz continuous function convex 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 $\frac{R^2B^2}{\epsilon^2}$ iterations.
- Advantages of this bound: a) Goes to zero as T gets large, and b) Independent of the dimensionality of $\mathbf{x}$!

How good or bad is this bound?

The analysis below assumes that we are always dealing with the same initial iterate x_1 (which determines R)

The recipe for number of iterates under this assumption is that if you want to get more close to the optimal, you would need number iterations inversely prportional to the square of how close you need to get to the optimal

Note that R is characterizing the initial iterate's distance only

- Final Result: Given a B-Lipschitz continuous function convex f, Gradient descent with step size $\gamma = \frac{R}{B\sqrt{T}}$ achieves a solution $\hat{x}$ s.t $|f(\hat{x}) - f(x^*)| \le \epsilon$ in $\frac{R^2 B^2}{\epsilon^2}$ iterations.
- Advantages of this bound: a) Goes to zero as T gets large, and b) Independent of the dimensionality of $\mathbf{x}$!

  Only the gradient computation will depend on the dimensionality of x

  The analysis is based on unit of each step which is a single gradient computaton
- Disadvantages: Slow convergence. To achieve a an error of 0.01, we require $10^4 R^2 B^2$ iterations. To achieve an error of 0.0001, the number of iterations is $10^8 R^2 B^2$!

Other disadvantages of the assumptions underlying this analysis of the algorithm

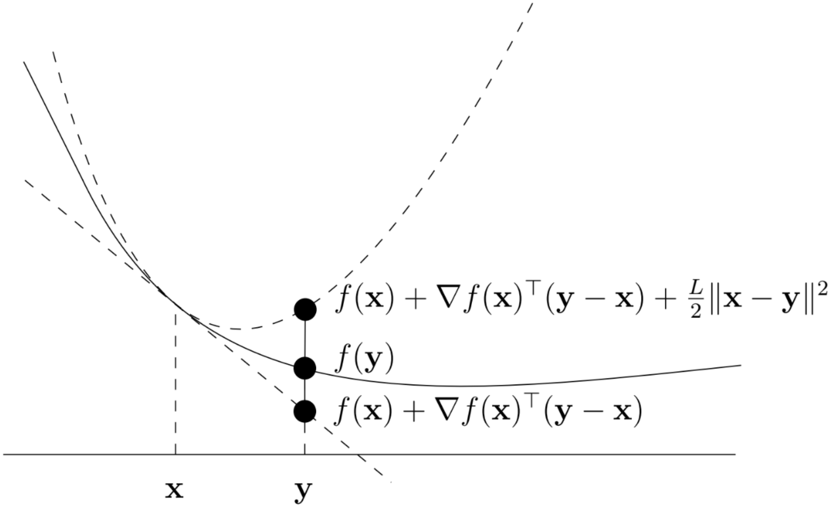
Does not assume that the step size γ is obtained in a more principled (search based) manner

See extra and optional slides: https://moodle.iitb.ac.in/pluginfile.php/143881/mod_resource/content/2/Optional%20Reading%20-%20Q-Convergence%20principles%2Bgeneral%20descent%20algorithms%2Bbacktracking%20ray%20search%2Barmijo%20conditions.pdf

Realistically gamma can be obtained using search techniques such as exact/backtracking ray search

Specifically backtracking ray search continue until conditions such as Armijo conditions/Goldstein conditions etc are satisfied

Can we do better using Lipschitz Smoothness of f?



Source: Martin Jaggi (CS 439)

Can we do better using Lipschitz Smoothness of f?

See the extra&optioonal slides at https://moodle.iitb.ac.in/pluginfile.php/143881/mod_resource/content/2/Optional%20Reading%20-%20Q-Convergence%20principles%2Bgeneral%20descent%20algorithms%2Bbacktracking%20ray%20search%2Barmijo%2

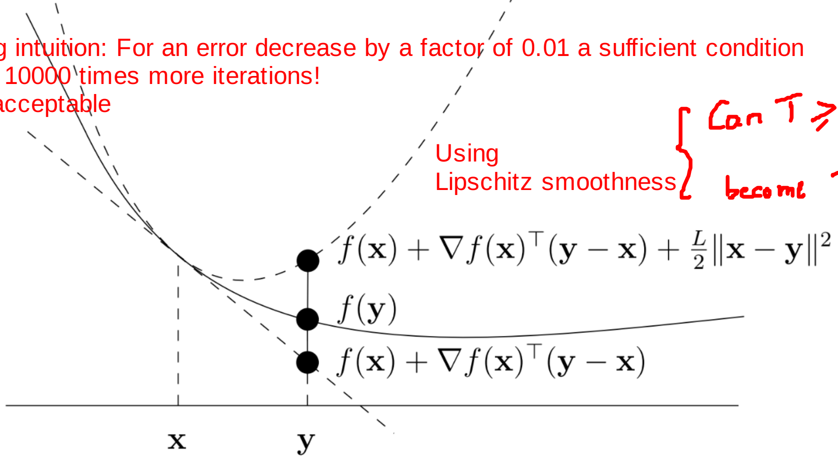
1) to give an idea of rate vs. order, Q-linear vs. Q-sublinear vs. Q-superlinear vs. R- convergence..
2) to help us appreciate "better" more rigorously

Our running intuition: For an error decrease by a factor of 0.01 a sufficient condition is need for 10000 times more iterations!
That is unacceptable

Using Lipschitz smoothness

Can $T \geq \frac{L^2 B^2}{\epsilon^2}$

become $T \geq \propto \frac{1}{\log(\epsilon)}$



Source: Martin Jaggi (CS 439)

Recap: Smoothness vs Continuity

- Bounded gradients $\iff$ Lipschitz continuous f
- Smoothness $\iff$ Lipschitz continuity of ∇f
- Properties of Lipschitz smoothness parameter L
  - Let $f_1, \cdots, f_m$ be smooth convex functions with parameters $L_1, \cdots, L_m$ and let $\lambda_1, \cdots, \lambda_m \geq 0$ be scalars. Then the convex function $f = \sum_{i=1}^{m} \lambda_i f_i$ is smooth with parameters $\sum_{i=1}^{m} \lambda_i L_i$
  - Let f be convex and smooth with parameter L and let $g(x) = Ax + b$ be a vector valued function. Then the convex function $f(g(x))$ is smooth with parameter $L||A||^2 = L\lambda_{\max}(A^T A)$. Here $||A||$ is the spectral norm of A.
  - Can you use this to derive a bound on the value of L for ∇f where f is the Logistic Loss? [Homework]
- Recall first order condition for Lipschitz smoothness:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$$

Recap: Smoothness vs Continuity

- Bounded gradients $\iff$ Lipschitz continuous f
- Smoothness $\iff$ Lipschitz continuity of ∇f
- Properties of Lipschitz smoothness parameter L
  - Let $f_1, \cdots, f_m$ be smooth convex functions with parameters $L_1, \cdots, L_m$ and let $\lambda_1, \cdots, \lambda_m \geq 0$ be scalars. Then the convex function $f = \sum_{i=1}^{m} \lambda_i f_i$ is smooth with parameters $\sum_{i=1}^{m} \lambda_i L_i$
  - Let f be convex and smooth with parameter L and let $g(x) = Ax + b$ be a vector valued function. Then the convex function $f(g(x))$ is smooth with parameter $L||A||^2 = L\lambda_{\max}(A^T A)$. Here $||A||$ is the spectral norm of A. $\rightarrow \lambda_{max} = ||A||_2 = \sup_{v \neq 0} \frac{||Av||_2}{||v||_2}$
  - Can you use this to derive a bound on the value of L for ∇f where f is the Logistic Loss? [Homework]
- Recall first order condition for Lipschitz smoothness:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$$

Recall: We have already discussed Lipschitz smoothness L of the Logistic Loss

Proof sketch: Using composition

① $f_1(t) = \log(1+e^t)$ $f_1'(t) = e^t/(1+e^t)$ $f_1''(t) = \nabla^2 f(t) = \frac{e^t(1+e^t) - e^t \cdot e^t}{(1+e^t)^2}$

$$= \frac{e^t}{(1+e^t)^2} \leq L \quad \{L = 1/4\}$$

Can be shown to be also Lipschitz Continuous

Using Quotient Rule

$$\frac{d\left(\frac{a(t)}{b(t)}\right)}{dt} = \frac{a'(t)b(t) - b'(t)a(t)}{(b(t))^2}$$

② $f_2(\theta) = -y_i \theta^T x_i$

THe Hessian Is indeed upper bounded by an L = 0

$\sum_i f_1(f_2(\theta))$ would therefore be Lipschizt Smooth

In fact it is also convex

Sketch of derivation

$$\sum f_1\big(\underbrace{f_2(\theta)}\big)$$

$$f_2(\theta) = -y_i x_i^T \theta = p_i^T \theta$$

$$\nabla f_1(f_2(\theta)) = f_1'(f_2(\theta))\, p_i$$

$$\propto \eta\, \theta^T \underbrace{p_i p_i^T}_{A_i} \le \eta \|\theta\| \, \|p_i^T p_i\|$$

By Cauchy Shwarz

$$L_{LR} = \eta \left(\frac{1}{4}\right) * \lambda_{max}(p_i p_i^T)$$

Gradient Descent for Smooth Functions: Analysis I

- Consider $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$

Gradient Descent for Smooth Functions: Analysis I

- Consider $f(\underset{\downarrow \; x_{t+1}}{\mathbf{y}}) \leq f(\underset{\downarrow \; x_t}{\mathbf{x}}) + \nabla^\top f(\mathbf{x})(\underbrace{\mathbf{y} - \mathbf{x}}_{-\gamma g_t}) + \frac{L}{2}\underbrace{||\mathbf{y} - \mathbf{x}||^2}_{\gamma^2 ||g_t||^2}$

Gradient Descent for Smooth Functions: Analysis I

- Consider $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$
- Note $x_{t+1} - x_t = -\gamma \nabla f(x_t) = -\gamma g_t$. Also substituting $y = x_{t+1}$ and $x = x_t$ above and doing some math, we obtain

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

$$\leq f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2 \tag{2}$$

Gradient Descent for Smooth Functions: Analysis I

- Consider $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}||\mathbf{y}-\mathbf{x}||^2$
- Note $x_{t+1} - x_t = -\gamma \nabla f(x_t) = -\gamma g_t$. Also substituting $y = x_{t+1}$ and $x = x_t$ above and doing some math, we obtain

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

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

Which trick to apply next?
Trick 1 (algebriac - expand in terms of squares)
Trick 2 (telescopic summing)
Trick 3 (minimize upper bound wrt parameters that do not characterize the LHS)

Gradient Descent for Smooth Functions: Analysis I

- Consider $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}||\mathbf{y}-\mathbf{x}||^2$
- Note $x_{t+1} - x_t = -\gamma \nabla f(x_t) = -\gamma g_t$. Also substituting $y = x_{t+1}$ and $x = x_t$ above and doing some math, we obtain

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

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

Which trick to apply next?
Trick 1 (algebriac - expand in terms of squares)
Trick 2 (telescopic summing)
Trick 3 (minimize upper bound wrt parameters that do not characterize the LHS)

Set $\frac{d\,RHS}{d\gamma} = 0$

i.e. $(-1 + L\gamma)||g_t||^2 = 0$

i.e. $\gamma = 1/L$

This value of gamma holds also in the worst case!

Gradient Descent for Smooth Functions: Analysis I

- Consider $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$
- Note $x_{t+1} - x_t = -\gamma \nabla f(x_t) = -\gamma g_t$. Also substituting $y = x_{t+1}$ and $x = x_t$ above and doing some math, we obtain

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

$$\leq f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2 \quad (2)$$

- **Minimizing upper bounds and maximizing lower bounds are frequently used tricks for convergence analysis since such an operation does not disrupt any inequality.** For what value of γ is $f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2$ minimized?

Gradient Descent for Smooth Functions: Analysis I

- Consider $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$
- Note $x_{t+1} - x_t = -\gamma \nabla f(x_t) = -\gamma g_t$. Also substituting $y = x_{t+1}$ and $x = x_t$ above and doing some math, we obtain

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

$$\leq f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2 \quad (2)$$

- **Minimizing upper bounds and maximizing lower bounds are frequently used tricks for convergence analysis since such an operation does not disrupt any inequality.** For what value of γ is $f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2$ minimized?
- Ans: For step size $\gamma = 1/L$. With this γ, the above result becomes:

Given the connection between spectral norm and L for smoothness, we see that more wobbly the function, less is the guaranteed decrease

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

Gradient Descent for Smooth Functions: Analysis I

- Consider $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}||\mathbf{y}-\mathbf{x}||^2$
- Note $x_{t+1} - x_t = -\gamma \nabla f(x_t) = -\gamma g_t$. Also substituting $y = x_{t+1}$ and $x = x_t$ above and doing some math, we obtain

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

$$\leq f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2 \quad (2)$$

- **Minimizing upper bounds and maximizing lower bounds are frequently used tricks for convergence analysis since such an operation does not disrupt any inequality.** For what value of γ is $f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2$ minimized?
- Ans: For step size $\gamma = 1/L$. With this γ, the above result becomes:

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

Even without convexity assumption L-smoothness gives some guaranteed decrease in every iteration!

Gradient Descent for Smooth Functions: Analysis I

- Consider $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}||\mathbf{y}-\mathbf{x}||^2$
- Note $x_{t+1} - x_t = -\gamma \nabla f(x_t) = -\gamma g_t$. Also substituting $y = x_{t+1}$ and $x = x_t$ above and doing some math, we obtain

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

$$\leq f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2 \quad (2)$$

- **Minimizing upper bounds and maximizing lower bounds are frequently used tricks for convergence analysis since such an operation does not disrupt any inequality.** For what value of γ is $f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2$ minimized?
- Ans: For step size $\gamma = 1/L$. With this γ, the above result becomes:

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

- This means GD is guaranteed to decrease the function value at every iteration!

Gradient Descent for Smooth Functions: Analysis I [Summarily]

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

- **Minimizing upper bounds and maximizing lower bounds are frequently used tricks for convergence analysis since such an operation does not disrupt any inequality.**
- For step size $\gamma = 1/L$, $f(x_t) - \gamma ||g_t||^2 + \frac{L}{2}\gamma^2 ||g_t||^2$ gets minimized. With this γ, the above result becomes:
$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L} ||g_t||^2$$
- This means GD is guaranteed to decrease the function value at every iteration!
- Recall that in the case of Lipschitz continuity, extreme right expression $\frac{\gamma}{2} TB^2 + \frac{R^2}{2\gamma}$ was minimized by setting its derivative to 0 which was obtained by setting $\gamma = \frac{R}{B\sqrt{T}}$.

Gradient Descent for Smooth Functions: Analysis I [Summarily]

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

- **Minimizing upper bounds and maximizing lower bounds are frequently used tricks for convergence analysis since such an operation does not disrupt any inequality.**
- For step size $\gamma = 1/L$, $f(x_t) - \gamma||g_t||^2 + \frac{L}{2}\gamma^2||g_t||^2$ gets minimized. With this γ, the above result becomes:

Note: These are value of gamma yielding the tightest upper bound

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

- This means GD is guaranteed to decrease the function value at every iteration!
- Recall that in the case of Lipschitz continuity, extreme right expression $\frac{\gamma}{2}TB^2 + \frac{R^2}{2\gamma}$ was minimized by setting its derivative to 0 which was obtained by setting $\gamma = \frac{R}{B\sqrt{T}}$.

Gradient Descent for Smooth Functions: Analysis I (concluded).

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

Gradient Descent for Smooth Functions: Analysis I (concluded).

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||g_t||^2 \Rightarrow \sum_{t=0}^{T-1} \frac{1}{2L}||g_t||^2 \leq \sum_{t=0}^{T-1} \left(f(x_t) - f(x_{t+1})\right)$$

Which trick/assumption to apply next?
Convexity assumption?
Telescopic summing?

Gradient Descent for Smooth Functions: Analysis I (concluded).

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

- Summing above inequality for $t = 0$ to $T - 1$:

$$\frac{1}{2L}\sum_{t=0}^{T-1}||g_t||^2 \leq \sum_{t=0}^{T-1}[f(x_t) - f(x_{t+1})] = [f(x_0) - f(x_T)] \tag{3}$$

Gradient Descent for Smooth Functions: Analysis I (concluded).

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

- Summing above inequality for $t = 0$ to $T-1$:

We have already got sufficient descrese from 0th to Tth iteration using L-smoothnes

$$\frac{1}{2L}\sum_{t=0}^{T-1}||g_t||^2 \leq \sum_{t=0}^{T-1}[f(x_t) - f(x_{t+1})] = [f(x_0) - f(x_T)] \qquad (3)$$

$$f(x_0) \cancel{- f(x_1) + f(x_1) - f(x_2)} \cdots - f(x_T)$$

CAN WE APPLY CONVEXITY ASSUMPTION NOW?

Gradient Descent for Smooth Functions: Analysis I (concluded).

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

- Summing above inequality for $t = 0$ to $T - 1$:

$$\frac{1}{2L}\sum_{t=0}^{T-1}||g_t||^2 \leq \sum_{t=0}^{T-1}[f(x_t) - f(x_{t+1})] = [f(x_0) - f(x_T)] \quad (3)$$

- Next, we present Analysis II, by invoking convexity (recall Analysis I & II from Gradient Descent for Lipschitz Continuity and Convex functions).

Gradient Descent for Smooth Functions: Analysis I (concluded).

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

- Summing above inequality for $t = 0$ to $T-1$:

$$\frac{1}{2L}\sum_{t=0}^{T-1}||g_t||^2 \le \sum_{t=0}^{T-1}[f(x_t) - f(x_{t+1})] = [f(x_0) - f(x_T)] \tag{3}$$

- Next, we present Analysis II, by invoking convexity (recall Analysis I & II from Gradient Descent for Lipschitz Continuity and Convex functions).

Recall: That in the previous analysis we firsty brought in Convexity and then used L-continuous (most convex functions are L-continuous)
Q: How to apply convexity such that x* (the optimal point) also starts appearing in the expressions?

Analysis II: Simple Expansion

- Define $g_t = \nabla f(x_t)$. From the definition of GD:

$$g_t^T(x_t - x^*) = \frac{1}{\gamma}(x_t - x_{t+1})^T(x_t - x^*)$$

- Note that $2v^Tw = ||v||^2 + ||w||^2 - ||v-w||^2$
- We can then rewrite the RHS as:

$$\begin{aligned} g_t^T(x_t - x^*) &= \frac{1}{2\gamma}(||x_t - x_{t+1}||^2 + ||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \\ &= \frac{\gamma}{2}||g_t||^2 + \frac{1}{2\gamma}(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \end{aligned} \tag{4}$$

- Summing (4) over $t = 0 \ldots T-1$ iterations:

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

Analysis II: Simple Expansion

- Define $g_t = \nabla f(x_t)$. From the definition of GD:

$$g_t^T(x_t - x^*) = \frac{1}{\gamma}(x_t - x_{t+1})^T(x_t - x^*)$$

- Note that $2v^Tw = ||v||^2 + ||w||^2 - ||v - w||^2$ Recall Trick 1
- We can then rewrite the RHS as:

$$\boxed{g_t^T(x_t - x^*)} = \frac{1}{2\gamma}(||x_t - x_{t+1}||^2 + ||x_t - x^*||^2 - ||x_{t+1} - x^*||^2)$$

$$= \boxed{\frac{\gamma}{2}||g_t||^2 + \frac{1}{2\gamma}(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2)} \tag{4}$$

- Summing (4) over $t = 0 \ldots T - 1$ iterations: Trick 2 again!

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

> 0

Analysis II: Invoking Convexity

- Invoking convexity with $x = x_t, y = x^*$.

$$f(x_t) - f(x^*) \leq g_t^T(x_t - x^*) \tag{5}$$

- Recall from (4) :

$$\boxed{\sum_{t=0}^{T-1} g_t^T(x_t - x^*) = \frac{1}{2\gamma}(||x_1 - x^*||^2 - ||x_T - x^*||^2) + \frac{\gamma}{2}\sum_{t=1}^{T-1}||g_t||^2}$$

which, based on $||x_T - x^*||^2 > 0$, implies:

$$\sum_{t=0}^{T-1} g_t^T(x_t - x^*) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2) \tag{6}$$

- Combining (5) with (6), we have:

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2 + \frac{1}{2\gamma}(||x_1 - x^*||^2)$$

Analysis II: Invoking Convexity

- Invoking convexity with $x = x_t, y = x^*$.

$$f(x_t) - f(x^*) \leq g_t^T(x_t - x^*) \quad (5)$$

- Recall from (4) :

$$\sum_{t=0}^{T-1} g_t^T(x_t - x^*) = \frac{1}{2\gamma}(||x_1 - x^*||^2 - ||x_T - x^*||^2) + \frac{\gamma}{2}\sum_{t=1}^{T-1}||g_t||^2$$

which, based on $||x_T - x^*||^2 > 0$, implies:

$$\sum_{t=0}^{T-1} g_t^T(x_t - x^*) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2) \quad (6)$$

- Combining (5) with (6), we have:

It can be shown that $\gamma = 1/L$ is good enough here as well to maintain the upper bound (since we had found the γ yielding lowest value of the RHS)

$$\boxed{\sum_{t=0}^{T-1}(f(x_t) - f(x^*))} \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2 + \frac{1}{2\gamma}(||x_1 - x^*||^2)$$

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

- The RHS, on setting $\gamma = 1/L$, yields

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

- Further, on invoking (3) on part of the RHS above

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

Analysis II: Invoking Convexity

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2 + \frac{1}{2\gamma}(||x_1 - x^*||^2)$$

- The RHS, on setting $\gamma = 1/L$, yields

It can be shown that $\Gamma = 1/L$ is good enough here as well to maintain the upper bound (since we had found the Γ yielding lowest value of the RHS)

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

- Further, on invoking (3) on part of the RHS above

$$\to \frac{1}{2L}\sum_{t=0}^{T-1}||g_t||^2 \leq f(x_0) - f(x_T$$

Ans: Take the terms in YELLOW to the LHS

$\cancel{f(x_0)}$

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

Q: How to go from here to the convergence...and speed of convergence...
That is, for upper bound on function value, number of iterations?

Gradient Descent for Smooth Functions: III

- We had:

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

Gradient Descent for Smooth Functions: III

- We had:

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

Ans: Take the terms in YELLOW
to the LHS

$\cancel{f(x_0)}$

$$\cancel{f(x_0)} - f(x^*) - \cancel{f(x_0)} + f(x_T) + \sum_{t=1}^{T-1} f(x_t) - f(x^*)$$

$$\sum_{t=1}^{T}[f(x_t) - f(x^*)]$$

Gradient Descent for Smooth Functions: III

- We had:

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

- Re-writing the math:

$$\sum_{t=1}^{T}(f(x_t) - f(x^*)) \leq \frac{L}{2}||x_0 - x^*||^2$$

Gradient Descent for Smooth Functions: III

- We had:

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

- Re-writing the math:

$$\sum_{t=1}^{T}(f(x_T) - f(x^*)) \leq \sum_{t=1}^{T}(f(x_t) - f(x^*)) \leq \frac{L}{2}||x_0 - x^*||^2 \leq \frac{LR^2}{2}$$

Recall: L-Smoothness guarantees that the function value descreases at every iteration

$\Rightarrow f(x_T) \leq f(x_t)$ for $t = 1 \ldots T-1$

Gradient Descent for Smooth Functions: III

- We had:

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

- Re-writing the math:

$$\sum_{t=1}^{T}(f(x_t) - f(x^*)) \leq \frac{L}{2}||x_0 - x^*||^2$$

- This implies that (**why?**):

$$f(x_T) - f(x^*) \leq \sum_{t=1}^{T}\frac{(f(x_t) - f(x^*))}{T} \leq \frac{L}{2T}||x_0 - x^*||^2$$

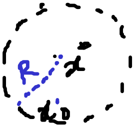
- We had:

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

- Re-writing the math:

$$\sum_{t=1}^{T}(f(x_t) - f(x^*)) \leq \frac{L}{2}||x_0 - x^*||^2$$

Assume



- This implies that (**why?**):

$$f(x_T) - f(x^*) \leq \sum_{t=1}^{T}\frac{(f(x_t) - f(x^*))}{T} \leq \frac{L}{2T}||x_0 - x^*||^2 = \frac{LR^2}{2T}$$

Recall: L-Smoothness guarantees that the function value descreases at every iteration

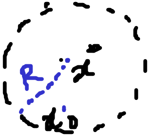
$\Rightarrow f(x_T) \leq f(x_t)$ for $t = 1 \ldots T-1$

Convergence rate for Smooth Functions

- Putting everything together: $f(x_T) - f(x^*) \leq \frac{L}{2T}||x_0 - x^*||^2 = \frac{LR^2}{2T}$

Convergence rate for Smooth Functions

Assume



- Putting everything together: $\underbrace{f(x_T) - f(x^*)}_{\leq \epsilon} \leq \frac{L}{2T}||x_0 - x^*||^2 = \underbrace{\frac{LR^2}{2T}}_{\leq \epsilon}$

For

Sufficient that

Convergence rate for Smooth Functions

- 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 $\frac{LR^2}{2T} \leq \epsilon$.

Convergence rate for Smooth Functions

- 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 $T \geq \frac{R^2 L}{2\epsilon}$

Convergence rate for Smooth Functions

- 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 $10^4 R^2 B^2$ in the Lipschitz case!

Convergence rate for Smooth Functions

- 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 $T \geq \frac{R^2L}{2\epsilon}$
- 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.

Convergence rate for Smooth Functions

- 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.

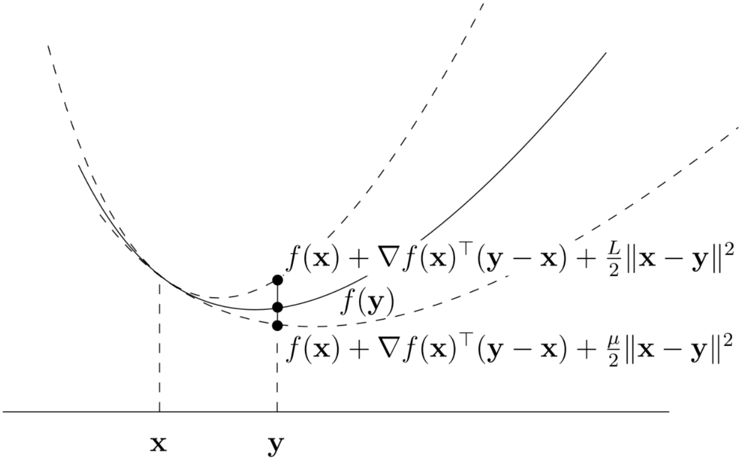
Recall this value was to give a lowest upper bound
In practice line/ray search techniques are used and convergence can be proved with Strong Wolfe conditions on step size

Pages 27-32 of https://moodle.iitb.ac.in/pluginfile.php/143881/mod_resource/content/2/Optional%20Reading%20-%20Q-Convergence%20p
Characterize Strong Wolfe condition

c1: Sufficient decrease of f with gamma

c2: Upper bounding increase of directional derivative of f with gamma

Smooth + Strongly Convex Functions



Source: Martin Jaggi (CS 439)

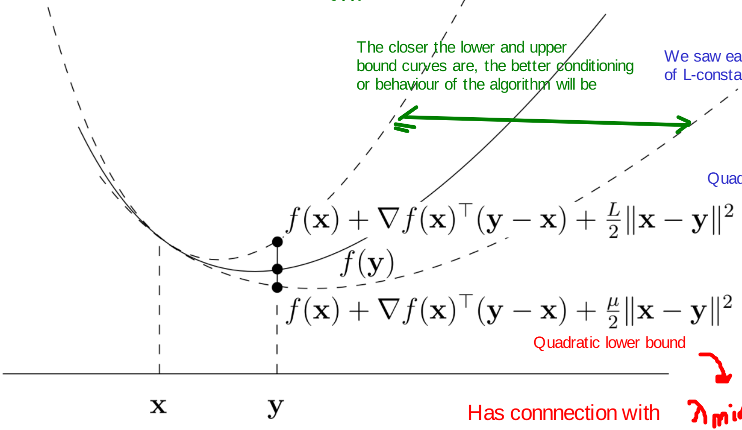
Smooth + Strongly Convex Functions

$\frac{\lambda_{max}}{\lambda_{min}}$: 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 Smooth + Strongly Convex I

- Recall from Analysis I:

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

Fastest Convergence with Smooth + Strongly Convex I

- Recall from Analysis I:

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

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

?