COM S 578X: Optimization for Machine Learning

Lecture Note 6: Gradient Descent

Jia (Kevin) Liu

Assistant Professor
Department of Computer Science
Iowa State University, Ames, Iowa, USA

Fall 2019

Outline

In this lecture:

- Convergence rate concept
- Gradient descent method
- Step size selection strategies
- Convergence performance of gradient descent

First-Order Algorithms: Smooth Convex Functions

Consider an unconstrained optimization problem, with f smooth and convex:

$A \preceq B \Leftrightarrow (B - A)$ is PSD.

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

twice cont. diff. $\rightarrow$ $\mathcal{F}_L^{2,1}$ $\leftarrow$ 1st der. UBed by L.

- Usually assume $\mu\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq L\mathbf{I}$, $\forall \mathbf{x}$, with $0 \le \mu \le L$ (using Nesterov's notation: $\mathcal{F}_L^{2,1}$, $\mathcal{S}_{\mu,L}^{2,1}$)

  twice cont. diff. $\rightarrow$ $\mathcal{S}_{\mu,L}^{2,1}$ $\leftarrow$ 1st der, UBed by L; strongly convex LBed by μ.

- If $\mu > 0$, then f is μ-strongly convex, i.e.,

  modulo

$$f(\mathbf{y}) \ge \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})}_{\text{FO approx.}} + \underbrace{\frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|_2^2}_{\ge 0}$$

- Condition number: $\kappa = L/\mu$ (the larger κ is, the more ill-conditioned)
- In ML, people are often interested in convex quadratics, e.g.,

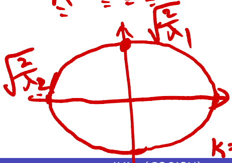
$A = U \Lambda U^\top$

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \underbrace{\mathbf{A}}_{\text{PSD.}}\mathbf{x}, \qquad \mu\mathbf{I} \preceq \mathbf{A} \preceq L\mathbf{I}$$

$= \frac{1}{2}(\underbrace{U^\top x}_{\tilde{x}})^\top \Lambda (\underbrace{U^\top x}_{\tilde{x}})$, $\tilde{x} \in \mathbb{R}^2$, $[\tilde{x}_1\ \tilde{x}_2]\begin{bmatrix}\frac{\lambda_1}{2} & \\ & \frac{\lambda_2}{2}\end{bmatrix}\begin{bmatrix}\tilde{x}_1\\ \tilde{x}_2\end{bmatrix} = \frac{\tilde{x}_1^2}{\frac{2}{\lambda_1}} + \frac{\tilde{x}_2^2}{\frac{2}{\lambda_2}} = 1$

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2, \qquad \mu\mathbf{I} \preceq \mathbf{A}^\top\mathbf{A} \preceq L\mathbf{I}$$

an ellipse



$\kappa = \frac{\lambda_1}{\lambda_2}$

Iterative Algorithms

We consider the following iterative algorithms:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + s_k \mathbf{d}_k,$$

where s_k is step-size, and $\mathbf{d}_k$ is search direction depending on $(\mathbf{x}_k, \mathbf{x}_{k-1}, \ldots)$.

For now: assume f smooth, $f(\mathbf{x}_k)$ and $\nabla f(\mathbf{x}_k)$ is easy to evaluate

Complications from ML:

- Nonsmooth f
- f not available (or too expensive to evaluate exactly)
- Only an estimate of $\nabla f(\mathbf{x}_k)$ is available
- A constraint $\mathbf{x} \in \Omega$ (usually a relatively simple Ω, e.g., ball, box, simplex...)
- Nonsmooth regularization, i.e., instead of $f(\mathbf{x})$, we want $\min f(\mathbf{x}) + \tau\psi(\mathbf{x})$

How to Evaluate the Speed of an Iterative Algorithm?

Definition 1 (Convergence rate)

A sequence $\{r_k\} \to r^*$ and $r_k \neq r^*$ for all k. The rate (or order) of convergence p is a nonnegative number satisfying

$$\limsup_{k\to\infty} \frac{\|r_{k+1} - r^*\|}{\|r_k - r^*\|^p} = \beta < \infty.$$

$c\frac{1}{k}$ iteration complexity.

ε

- Sublinear: $p = 1$ and $\beta = 1$ (e.g., $O(1/k)$ rate) i.e., $\frac{\|r_{k+1} - r^*\|}{\|r_k - r^*\|} \to 1$, as $k \to \infty$

  $\frac{\frac{c}{k+1}}{\frac{c}{k}} = \frac{k}{k+1} \to 1$ as $k \to \infty$. Desired ε: $\frac{c}{k} \le \varepsilon \Rightarrow \boxed{k \ge \frac{c}{\varepsilon} = \Omega\left(\frac{1}{\varepsilon}\right)}$

  contraction mapping

- Linear or geometric: $p = 1$ and $0 < \beta < 1$ (i.e., $\|r_{k+1} - r^*\| \le \beta\|r_k - r^*\|$ for some $\beta \in (0, 1)$, or $\|r_k - r^*\| = O(\beta^k)$, which is quite fast) $\le \beta^2 \|r_{k-1} - r^*\| \cdots \le \beta^k \underbrace{\|r_0 - r^*\|}_{\text{const}} = O(\beta^k)$.

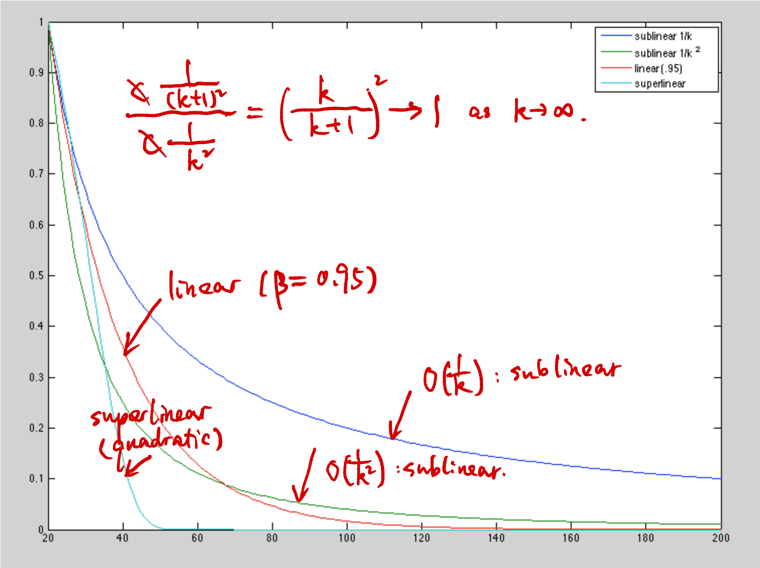
  Desired ε: $c\beta^k \le \varepsilon \Rightarrow k \ge c' \log\left(\frac{1}{\varepsilon}\right)$: Need $O\left(\log\frac{1}{\varepsilon}\right)$ iter.

- Superlinear: $p > 1$ and $\beta < \infty$, or $p = 1$ and $\beta = 0$ (i.e., $\frac{\|r_{k+1}-r^*\|}{\|r_k - r^*\|} \to 0$, that's very fast!) Not only a contraction mapping but also, the rate of contraction is accelerating!

- Quadratic: $p = 2$ and $\beta < \infty$ (i.e., $\|r_{k+1} - r^*\| \le \beta\|r_k - r^*\|^2$, # of correct significant digits doubles each iteration. We rarely need anything faster than this!) For ε-accuracy: Need $O\left(\log\log\left(\frac{1}{\varepsilon}\right)\right)$ iterations $\leftarrow$ almost const.

p, β

Convergence Rates Comparisons



$$\frac{\propto \frac{1}{(k+1)^2}}{\propto \frac{1}{k^2}} = \left(\frac{k}{k+1}\right)^2 \to 1 \text{ as } k \to \infty.$$

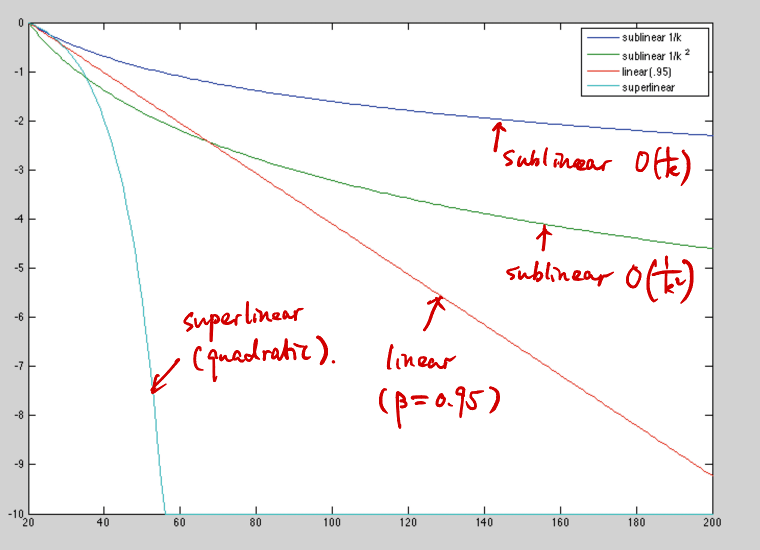
linear ($\beta = 0.95$)

$O(\frac{1}{k})$: sublinear

superlinear (quadratic)

$O(\frac{1}{k^2})$: sublinear.

Convergence Rates Comparisons: Log-Scale



Gradient Descent

Back to the unconstrained optimization problem, with f smooth and convex:

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

Denote the optimal value as $f^* = \min_{\mathbf{x}} f(\mathbf{x}^*)$ and an optimal solution as $\mathbf{x}^*$

Gradient Descent

Choose initial point $\mathbf{x}_0 \in \mathbb{R}^n$. Repeat:

$$\mathbf{x}_k = \mathbf{x}_{k-1} - s_k \nabla f(\mathbf{x}_{k-1}), \quad k = 1, 2, 3, \dots$$

Stop if some stopping criterion is satisfied. (e.g., $\|\nabla f(\underline{x}_k)\| \le \varepsilon$, some # of iter, ...)

Gradient Descent: Geometric Interpretation

Gradient descent is a first-order method: Consider the following quadratic Taylor approximation:

$$f(\mathbf{y}) \approx \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x})}_{\text{FO approx.}} + \underbrace{\frac{1}{2}(\mathbf{y}-\mathbf{x})^\top \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})} \cancel{+ o(\|\mathbf{y}-\mathbf{x}\|^2)}$$

$\frac{1}{s}\mathbf{I}$

No, we replace Hessian $\nabla^2 f(\mathbf{x})$ by $\frac{1}{s}\mathbf{I}$ to obtain:

$$f(\mathbf{y}) \approx \underbrace{f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) \boxed{+ \frac{1}{2s}\|\mathbf{y}-\mathbf{x}\|^2}}_{F(\mathbf{y})}$$

"proximity term": penalize moving too far from $\mathbf{x}$

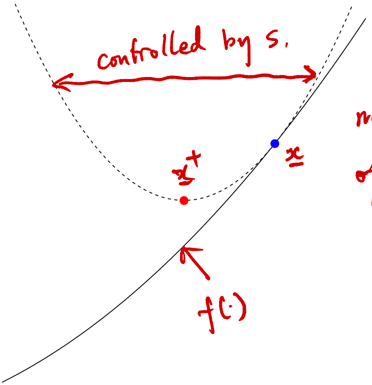
Can be viewed as a linear approximation to f, with proximity term to $\mathbf{x}$ weighted by $\frac{1}{2s}$. Choose next point $\mathbf{y} = \mathbf{x}^+$ to minimize this approximation:

$$\mathbf{x}^+ = \mathbf{x} - s\nabla f(\mathbf{x})$$

Quad fn of y, unconstr: Set grad to 0, then solve for y.

$\nabla F(\mathbf{y}) = \mathbf{0} \Rightarrow \nabla f(\mathbf{x}) + \frac{1}{s}(\mathbf{y}-\mathbf{x}) = \mathbf{0} \Rightarrow \mathbf{y} = \mathbf{x} - s\nabla f(\mathbf{x})$.

Gradient Descent: Geometric Interpretation



smaller s : narrower the openning of parabola.

$$\mathbf{x}^+ = \arg\min_y f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{1}{2s}\|\mathbf{y}-x\|_2^2$$

Questions:

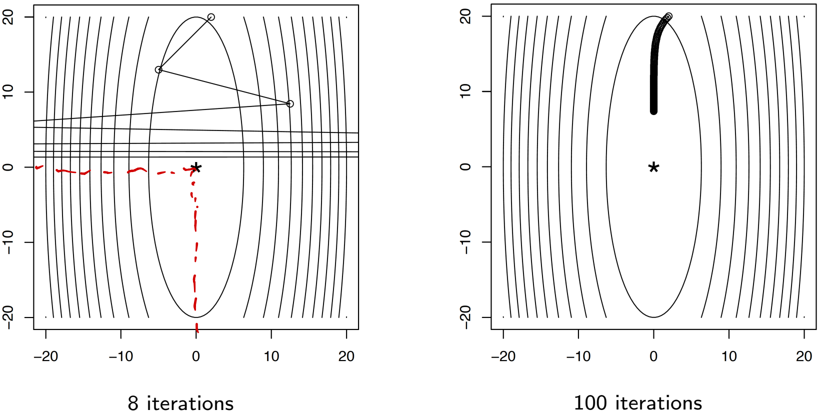
- How to choose step sizes $\{s_k\}$?
- What is the according convergence rate? Or does it depend on $\{s_k\}$? Yes!

Strategy 1: Fixed Step Size

Simply set $s_k = s$ for all $k = 1, 2, 3, \ldots$.

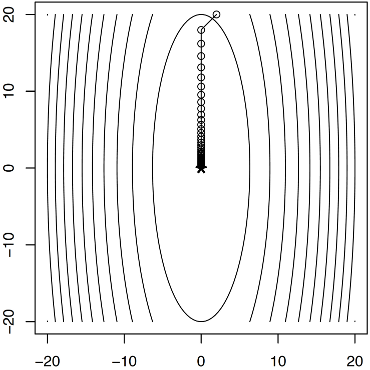
Limitations: May diverge if s is too large, Can be slow if s is too small.

Example: Consider $f(\mathbf{x}) = (10x_1^2 + x_2^2)/2$: $\Rightarrow (x_1^*, x_2^*) = (0, 0)$.



Strategy 1: Fixed Step Size

Converges nicely when s is "just right." Same example, GD after 40 iterations:



Will be clear what we mean by "just right" in convergence rate analysis later

Need info of the "Lipschitz const." of $\nabla f(\underline{x})$.

Strategy 2: Exact Line Search

Choose the step size s to do the "best" we can along the direction of $-\nabla f(\mathbf{x})$:

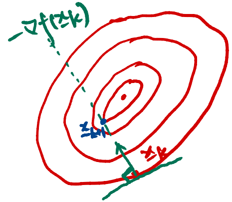
$$s = \arg\min_{t \geq 0} f(\mathbf{x} - t\nabla f(\mathbf{x}))$$

Take directional der., set it to 0, and solve for t:

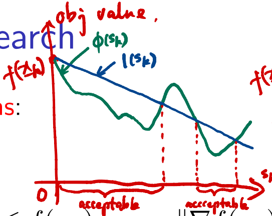
$$-\nabla f(x_{k+1})^T \nabla f(x_k) = -\nabla f(x_k - t\nabla f(x_k))^T \nabla f(x_k) = 0$$

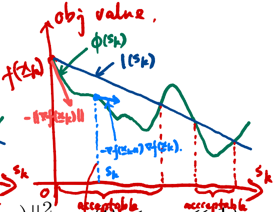
Limitations:

- Usually it's too expensive to do this in each iteration.
- Spoiler: Our convergence rate analysis later will also show that it's not worth the effort



Strategy 3: Inexact Line Search





Seek s_k that satisfies Wolfe conditions:

- "Sufficient decrease" in f:

$$f(\mathbf{x}_{k+1}) = \underbrace{f(\mathbf{x}_k - s_k \nabla f(\mathbf{x}_k))}_{/\!/\ \phi(s_k)} \le \underbrace{f(\mathbf{x}_k) - c_1 s_k \|\nabla f(\mathbf{x}_k)\|^2}_{= l(s_k)}, \quad (0 < c_1 \ll 1)$$

if $\phi'(s_k)$ is not too neg. (or even pos.), then we should STOP!

- "Not zigzagging too badly":

$$\underbrace{-\nabla f(\mathbf{x}_{k+1})^\top \nabla f(\mathbf{x}_k)}_{\text{directional der. of } \phi(s_k)} \ge -c_2 \|\nabla f(\mathbf{x}_k)\|^2, \quad (c_1 < c_2 < 1)$$

Main features: w.r.t. s_k (i.e., $\phi'(s_k)$) : $\phi'(s_k) = -\nabla f(\underline{x}_{k+1})^\top \nabla f(\underline{x}_k)$

- Can show that accumulation points $\bar{\mathbf{x}}$ of $\{\mathbf{x}_k\}$ are stationary: $\nabla f(\bar{\mathbf{x}})$ (thus minimizer if f is convex)
- Can do 1-dim line search for s_k, taking minima of quadratic or cubic interpolations of f and ∇f at the last two values tried. Use brackets for reliability. Often finds suitable s_k within 3 attempts (see [Nocedal & Wright, 2006, Ch. 3])

Strategy 3: Inexact Line Search – Backtracking

One way to adaptively choose step size is to use backtracking line search

1. First fix parameters $0 < \beta < 1$ and $0 < \alpha \leq \frac{1}{2}$
2. At each iteration, start with $s = 1$, and while

or start w/ s_{int}

$$f(\mathbf{x} - s\nabla f(\mathbf{x})) > f(\mathbf{x}) - \alpha s \|\nabla f(\mathbf{x})\|_2^2$$

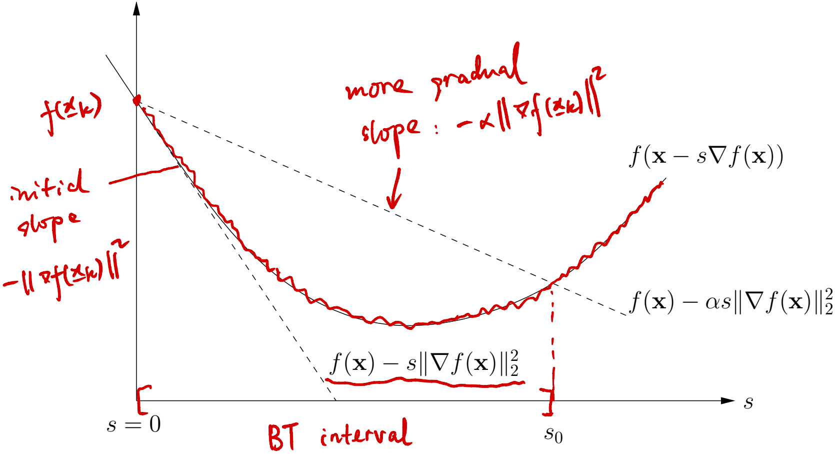
shink $\underbrace{s = \beta s}_{\text{shrink by a factor } \beta}$. Else, perform gradient descent update:

$$\mathbf{x}^+ = \mathbf{x} - s\nabla f(\mathbf{x})$$

Remarks:

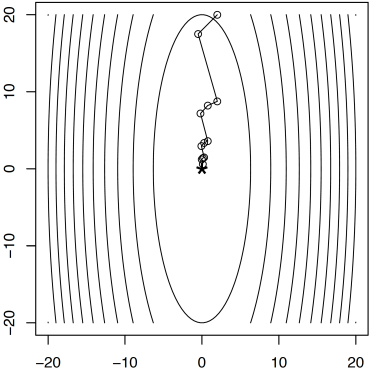
- Simple and tends to work well in practice (further simplification: just take $\alpha = \beta = 1/2$). But doesn't work for f nonsmooth
- Also referred to as Armijo's rule. Step size shrinking very aggressively
- Not checking the second Wolfe condition: the s_k thus identified is "within striking distance" of an s that's not too large

Backtracking Interpretation



Backtracking Example

Backtracking picks up roughly the right step size (12 outer iterations, 40 iterations in total):



Convergence Rate Analysis: Fixed Step Size

Assume that f is convex & differentiable, with $\mathrm{dom}(f) = \mathbb{R}^n$ and additionally

$$\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\|_2 \leq L\|\mathbf{y} - \mathbf{x}\|_2, \quad \forall \mathbf{x}, \mathbf{y}$$

That is, ∇f is Lipschitz continuous with constant $L > 0$ (L-Lipschitz continuous) (Nesterov notation: $f \in \mathcal{F}_L^{1,1}$)

Theorem 1

Gradient descent with fixed step size $s \leq 1/L$ satisfies

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2sk}, = O\left(\frac{1}{k}\right).$$

i.e., gradient descent method has sublinear convergence rate $O(1/k)$.

Remark:

- To get $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$, it takes $O(1/\epsilon)$ iterations.

Convergence Rate Analysis: Fixed Step Size

Proof Sketch.

- ∇f is L-Lipschitz $\Rightarrow$

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \quad \forall \mathbf{x}, \mathbf{y}$$

- Plugging in $\mathbf{x}_{k+1} = \mathbf{x}_k - s\nabla f(\mathbf{x}_k)$ to obtain:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \left(1 - \frac{Ls}{2}\right) s\|\nabla f(\mathbf{x}_k)\|_2^2$$

- Using the convexity of f and taking $0 < s \leq 1/L$, and , we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}^*) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_k - \mathbf{x}^*) - \frac{s}{2}\|\nabla f(\mathbf{x}_k)\|_2^2 \\ &= f(\mathbf{x}^*) + \frac{1}{2s}\left(\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2\right) \end{aligned}$$

Convergence Rate Analysis: Fixed Step Size

- Summing over iterations & after telescoping:

$$\begin{aligned}\sum_{i=1}^{k}\big(f(\mathbf{x}_i) - f(\mathbf{x}^*)\big) &\le \frac{1}{2s}\big(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_k - \mathbf{x}^*\|_2^2\big) \\ &\le \frac{1}{2s}\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2\end{aligned}$$

- Since $f(\mathbf{x}_k)$ is non-increasing, we have

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \frac{1}{k}\sum_{i=1}^{k}\big(f(\mathbf{x}_i) - f(\mathbf{x}^*)\big) \le \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2sk}. \qquad \square$$

GD Convergence: Fixed Step Size under Strong Convexity

Assume that f is convex is differentiable, ∇f L-Lipschitz, and μ-strongly convex, i.e., $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2}\mu \|\mathbf{y} - \mathbf{x}\|_2^2$. (Nesterov notation: $\mathcal{S}_{\mu,L}^{1,1}$)

Theorem 2

Gradient descent with fixed step size $0 < s \leq 2/(L+\mu)$ satisfies

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \left(\sqrt{1 - \frac{2s\mu L}{\mu + L}}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|$$

$$\sqrt{1-\frac{2}{\mu+L}\times\frac{2}{\mu+L}}$$

i.e., GD has linear convergence rate. If $s = \frac{2}{\mu+L}$, then

$$= \sqrt{1-\frac{4}{(\mu+L)^2}} = \sqrt{\frac{(L-\mu)^2}{(L+\mu)^2}}$$

$$= \frac{L-\mu}{L+\mu} = \frac{\kappa-1}{\kappa+1}$$

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^k \|\mathbf{x}_0 - \mathbf{x}^*\|$$

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{L}{2}\left[\left(\frac{\kappa - 1}{\kappa + 1}\right)^2\right]^k \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2$$

$=0$

$$f(x_k) - f(x^*) \leq \cancel{\nabla f(x^*)(y-x)} + \frac{L}{2}\|y-x^*\|^2 \leq \frac{L}{2}\left(\frac{\kappa-1}{\kappa+1}\right)^{2k}\|x_0-x^*\|_2^2$$

GD Convergence: Fixed Step Size under Strong Convexity

Proof Sketch. For notational convenience, let r_k denote the residual $\|\mathbf{x}_k - \mathbf{x}^*\|$.

- Consider r_{k+1}^2, we have

$$\begin{aligned} r_{k+1}^2 &= \|\mathbf{x}_{k+1} - \mathbf{x}^*\|^2 = \|\mathbf{x}_k - \mathbf{x}^* - s\nabla f(\mathbf{x}_k)\|_2^2 \\ &= r_k^2 - 2s\nabla f(\mathbf{x}_k)^\top(\mathbf{x}_k - \mathbf{x}^*) + s^2\|\nabla f(\mathbf{x}_k)\|_2^2 \end{aligned} \tag{1}$$

- According to [Nesterov, Thm 2.1.12], if $f \in \mathcal{S}_{\mu,L}^{1,1}$, we have

$$\nabla f(\mathbf{x}_k)^\top(\mathbf{x}_k - \mathbf{x}^*) \geq \frac{\mu L}{\mu + L} r_k^2 + \frac{1}{\mu + L}\|\nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}^*)\|_2^2$$

- Plugging in (1) and using the fact that $\nabla f(\mathbf{x}^*) = 0$, we have:

$$r_{k+1}^2 \leq \left(1 - \frac{2s\mu L}{\mu + L}\right) r_k^2 + s\left(s - \frac{2}{\mu + L}\right)\|\nabla f(\mathbf{x}_k)\|_2^2$$

- The last inequality in Thm 2 follows from the L-Lipschitz gradient assumption. □

Convergence Rate Analysis: Backtracking

Same assumption: $f \in \mathcal{F}_L^{1,1}$.

Theorem 3

Gradient descent with backtracking line search satisfies:

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{4\underbrace{\alpha \min\{1, \beta/L\}}_{\hat{s}/2} k} = O(1/k)$$

$$\frac{\|\underline{x}_0 - \underline{x}^*\|^2}{2\hat{s}k}$$

- Same sublinear rate as fixed step size
- If β is not too small, then we don't lose much compared to fixed step size (β/L vs $1/L$)

Convergence Rate Analysis: Backtracking

$0 < \alpha \le \frac{1}{2}$

Proof.

- Recall BT exit condition: $f(\mathbf{x}_k - s\nabla f(\mathbf{x}_k)) \le f(\mathbf{x}_k) - \alpha s \|\nabla f(\mathbf{x}_k)\|_2^2$. This is satisfied when $s \le 1/L$, because $s \le 1/L \Rightarrow -s + \frac{Ls^2}{2} \le -\frac{s}{2}$
- Using this and the L-Lipschitz assumption, we have

$0 < \alpha \le \frac{1}{2}$

$$f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k - s\nabla f(\mathbf{x}_k)) \le f(\mathbf{x}_k) - \frac{s}{2}\|\nabla f(\mathbf{x})\|_2^2 \le f(\mathbf{x}) - \alpha s \|\nabla f(\mathbf{x}_k)\|_2^2$$

In const. step:

$f(x_{k+1}) \le f(x_k) - s(1 - \frac{Ls}{2}) \|\nabla f(x_k)\|^2$

$\beta \in (0,1)$.

- Hence, BTLS terminates either with $s = 1$ or with $s \ge \beta/L$. Thus, we have

$$f(\mathbf{x}_{k+1}) \le \underline{f(\mathbf{x}_k)} - \min\{\alpha, \beta\alpha/L\}\|\nabla f(\mathbf{x}_k)\|_2^2$$

$\le f(x^*) + \nabla f(x_k)^T (x_k - x^*) - \boxed{\min\{\alpha, \frac{\beta\alpha}{L}\}} \|\nabla f(x_k)\|_2^2 \quad (\Delta)$

- The rest of the proof follows essentially the same line of arguments as in the proof for the fixed step size case. □

Call: $\frac{\hat{s}}{2}$

(starting from (Δ), the rest just follows from fixed step-size case).

GD Convergence: Backtracking under Strong Convexity

Assume that $f \in \mathcal{S}^{1,1}_{\mu,L}$

Theorem 4

Gradient descent with backtracking line search satisfies

$$\|f(\mathbf{x}_k) - f(\mathbf{x}^*)\| \leq \left(1 - \min\{2\mu\alpha, 2\beta\alpha\mu/L\}\right)^k \|f(\mathbf{x}_0) - f(\mathbf{x}^*)\|$$

i.e., GD has a linear convergence rate.

GD Convergence: Backtracking under Strong Convexity

Proof.

- From the proof of the weakly convex case, we have obtained:

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \min\{\alpha, \beta\alpha/L\} \|\nabla f(\mathbf{x}_k)\|_2^2$$

$-f(x^*)$ $-f(x^*)$

- Noting $\|\nabla f(\mathbf{x}_k)\|_2^2 \geq 2\mu(f(\mathbf{x}_k) - f(\mathbf{x}^*))$ & subtracting $f(\mathbf{x}^*)$ on both sides:

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq (1 - \min\{2\mu\alpha, 2\beta\alpha\mu/L\}) (f(\mathbf{x}_k) - f(\mathbf{x}^*)).$$

$\square$

Proof. If $f \in S^{1,1}_{\mu,L}$, recall we've shown:

$f(y) \geq f(x) + \nabla f(x)^T(y - x) + \frac{\mu}{2}\|y - x\|^2$ (def. of strong convexity).

RHS is convex & quad fn. of y (given fixed x).

Minimize RHS w.r.t. y (find least LB) $\Rightarrow \tilde{y} = x - \frac{1}{\mu}\nabla f(x)$. (Δ)

Plug (Δ) into RHS yields: $f(y) \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2$. Multiply both sides by 2μ rearranging. $\boxtimes$

Convergence Rate Analysis: Exact LS

Assume that $f \in \mathcal{F}_L^{1,1}$

Theorem 5

Gradient descent with exact line search satisfies

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{L\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2}{2k} = O(1/k)$$

i.e., GD has a sublinear convergence rate.

Convergence Rate Analysis: Exact LS

Proof.

- From L-Lipschitz and $\mathbf{x}_{k+1} = \mathbf{x}_k - s\nabla f(\mathbf{x}_k)$, we have

optimal step-size is also UBed by the min of the RHS

$$f(\mathbf{x}_k - s\nabla f(\mathbf{x}_k)) \leq \underline{f(\mathbf{x}) - s\left(1 - \frac{Ls}{2}\right)\|\nabla f(\mathbf{x}_k)\|_2^2}$$

$$g(s) = f(x) + \frac{L}{2}\|\nabla f(x_k)\|_2^2 s^2 - s\|\nabla f(x_k)\|_2^2$$

- Minimize over s on both sides yields:

$$\Rightarrow s_e = \frac{1}{L}.$$

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k - {\color{red}s_e}\nabla f(\mathbf{x}_k)) \leq \underline{f(\mathbf{x}_k)} - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2$$

$$\leq f(x^*) + \nabla f(x_k)^T(x_k - x^*) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2$$

$\hat{s}$

- The rest of the proof follows exactly the same arguments as in the proof of the fixed step size case □

(DIY)

GD Convergence: Exact LS under Strong Convexity

Assume that $f \in \mathcal{S}^{2,1}_{\mu,L}$

Theorem 6

Gradient descent with exact line search satisfies

$$\|f(\mathbf{x}_k) - f(\mathbf{x}^*)\| \leq (1 - \mu/L)^k \|f(\mathbf{x}_0) - f(\mathbf{x}^*)\|$$

i.e., GD has a linear convergence rate.

Observation

No improvement in the linear rate over fixed step size!

GD Convergence: Exact LS under Strong Convexity

Proof.

- From L-Lipschitz and $\mathbf{x}_{k+1} = \mathbf{x}_k - s\nabla f(\mathbf{x}_k)$, we have

$$f(\mathbf{x}_k - s\nabla f(\mathbf{x}_k)) \leq f(\mathbf{x}) - s\left(1 - \frac{Ls}{2}\right)\|\nabla f(\mathbf{x}_k)\|_2^2$$

- Minimize over s on both sides yields:

$-f(x^*)$ $\qquad\qquad$ $-f(x^*)$

$$f(\mathbf{x}_{k+1}) = f(\mathbf{x}_k - s_e\nabla f(\mathbf{x}_k)) \leq f(\mathbf{x}_k) - \frac{1}{2L}\|\nabla f(\mathbf{x}_k)\|_2^2$$

- Following the same step in BTLS, subtracting $f(\mathbf{x}^*)$ from both sides and noting Noting $\|\nabla f(\mathbf{x}_k)\|_2^2 \geq 2\mu(f(\mathbf{x}_k) - f(\mathbf{x}^*))$, we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq (1 - \underbrace{\mu/L}_{1/\kappa})\,(f(\mathbf{x}_k) - f(\mathbf{x}^*)).$$

□

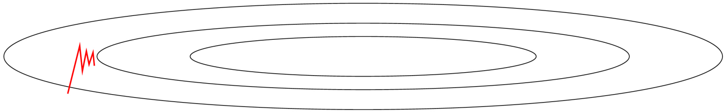
In Summary

	Convex	Strongly Convex
Fixed Step Size	$O(1/k)$	$O\left[\left(\frac{\kappa-1}{\kappa+1}\right)^{2k}\right]$
BTLS	$O(1/k)$	$O((1-\min\{2\mu\alpha, 2\beta\alpha/\kappa\})^k)$
Exact LS	$O(1/k)$	$O\left[\left(\frac{\kappa-1}{\kappa}\right)^{k}\right]$

only slightly better.

The Slow Linear Rate Is Typical

Not just pessimistic bound – It really is quite slow!



Next Class

Accelerated First-Order Methods

Thm 1: If $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^n)$, then GD with const. step-size $s \in (0, \frac{1}{L}]$ satisfies:

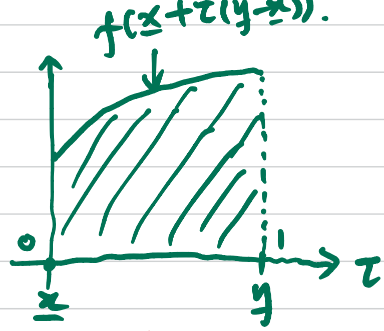
$$f(x_k) - f(x^*) \le \frac{\|x_0 - x^*\|^2}{2sk}, \ \forall k$$

($O(\frac{1}{k})$ sublinear convergence rate).

Proof. step ① Claim: If ∇f is Lipschitz, then:

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|_2^2, \ \forall x, y \in \mathbb{R}^n. \quad (\star)$$

To show $(\star)$, we start from the following:



$$f(y) = f(x) + \int_0^1 \underbrace{f'(x+\tau(y-x))}_{} d\tau$$

dir. der.

$$= f(x) + \int_0^1 \nabla f(x+\tau(y-x))^T (y-x)\, d\tau \qquad \text{(chain rule)}$$

add & subtract

$$= f(x) + \nabla f(x)^T (y-x) + \int_0^1 [\nabla f(x+\tau(y-x)) - \nabla f(x)]^T (y-x)\, d\tau$$

By some rearranging and taking absolute value on both sides:

$$|f(y) - f(x) - \nabla f(x)^T (y-x)| = \left| \int_0^1 [\nabla f(x+\tau(y-x)) - \nabla f(x)]^T (y-x)\, d\tau \right|$$

$$\le \int_0^1 |[\nabla f(x+\tau(y-x)) - \nabla f(x)]^T (y-x)|\, d\tau. \qquad \begin{pmatrix}\text{Triangle Ineq.} \\ \|a+b\| \le \|a\| + \|b\|\end{pmatrix}$$

$$\le \int_0^1 \underbrace{\|\nabla f(x+\tau(y-x)) - \nabla f(x)\|}_{L\text{-Lipschitz}:\ \le L\tau\|y-x\|.} \cdot \|y-x\|\, d\tau \qquad \begin{pmatrix}\text{Cauchy-Schwatz Ineq.} \\ |a^T b| \le \|a\| \cdot \|b\|\end{pmatrix}$$

$$\le \int_0^1 L\tau \|y-x\|^2 d\tau = L\|y-x\|^2 \underbrace{\int_0^1 \tau\, d\tau}_{\frac{1}{2}} = \frac{L}{2}\|y-x\|^2. \ (\star) \text{ is proved.}$$

Step ②: WTS: "Descent property of GD":

$x_{k+1} = x_k - s_k \nabla f(x_k)$. Plugg this in (☆):

$$f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T \overbrace{(x_{k+1} - x_k)}^{-s\nabla f(x_k)} + \frac{L}{2} \| \overbrace{x_{k+1} - x_k}^{-s\nabla f(x_k)} \|^2$$

$$= f(x_k) - s \|\nabla f(x_k)\|_2^2 + \frac{Ls^2}{2} \|\nabla f(x_k)\|_2^2$$

$$= f(x_k) - s\left(1 - \frac{Ls}{2}\right) \|\nabla f(x_k)\|_2^2 \qquad (\Delta)$$

Step ③: From convexity of $f(x)$, we have:

$$f(x^*) \ge f(x_k) + \nabla f(x_k)^T (x^* - x_k)$$

$$\Rightarrow f(x_k) \le f(x^*) + \nabla f(x_k)^T (x_k - x^*) \qquad (*)$$

Plugging $(*)$ into (Δ) yields:

$$f(x_{k+1}) \le \underbrace{f(x^*) + \nabla f(x_k)^T (y - x)}_{\text{replace } f(x_k)} - s\left(1 - \frac{Ls}{2}\right) \|\nabla f(x_k)\|_2^2 \qquad (**)$$

Now, let's take step-size $s \in (0, \frac{1}{L}]$, then

$$0 < s \le \frac{1}{L} \Rightarrow 0 \le Ls \le 1 \Rightarrow -\frac{1}{2} \le -\frac{Ls}{2} < 0 \Rightarrow \frac{1}{2} \le 1 - \frac{Ls}{2} < 1$$

$$\Rightarrow -s < -s\left(1 - \frac{Ls}{2}\right) \le -\frac{s}{2}$$

Using above in $(**)$ $\Rightarrow$

$$f(x_{k+1}) - f(x^*) \le \underline{\nabla f(x_k)^T (x_k - x^*) - \frac{s}{2} \|\nabla f(x_k)\|_2^2} \qquad (\Delta\Delta).$$

Step ④: Consider RHS: $\nabla f(\underline{x}_k)^T(\underline{x}_k-\underline{x}^*)-\frac{s}{2}\|\nabla f(\underline{x}_k)\|_2^2$

$$\nabla f(\underline{x}_k)^T(\underline{x}_k-\underline{x}^*)-\frac{s}{2}\|\nabla f(\underline{x}_k)\|_2^2$$

$$= -\frac{1}{2s}\left[\underbrace{s^2\|\nabla f(\underline{x}_k)\|_2^2 - 2s\nabla f(\underline{x}_k)^T(\underline{x}_k-\underline{x}^*)+\|\underline{x}_k-\underline{x}^*\|_2^2}-\|\underline{x}_k-\underline{x}^*\|_2^2\right]$$

$$= -\frac{1}{2s}\left[\|(\underline{x}_k-\underline{x}^*)-s\nabla f(\underline{x}_k)\|_2^2-\|\underline{x}_k-\underline{x}^*\|_2^2\right]$$

$$= \frac{1}{2s}\left[\|\underline{x}_k-\underline{x}^*\|_2^2-\|\underline{x}_{k+1}-\underline{x}^*\|_2^2\right]$$

Therefore, $(\Delta\Delta) \Rightarrow$

$$f(\underline{x}_{k+1})-f(\underline{x}^*) \le \frac{1}{2s}\left(\|\underline{x}_k-\underline{x}^*\|_2^2-\|\underline{x}_{k+1}-\underline{x}^*\|_2^2\right) \qquad (\square)$$

Step ⑤: Summing $(\square)$ from 1 to k (telescoping):

$$\sum_{i=1}^{k}\left(f(\underline{x}_i)-f(\underline{x}^*)\right) \le \frac{1}{2s}\left(\|\underline{x}_0-\underline{x}^*\|_2^2-\|\underline{x}_k-\underline{x}^*\|_2^2\right)$$

$$\le \frac{1}{2s}\|\underline{x}_0-\underline{x}^*\|_2^2$$

Since $\{f(\underline{x}_k)\}$ is mono. non-incr. (GD descent prop.), we have

$$f(\underline{x}_k)-f(\underline{x}^*) \le \frac{1}{k}\sum_{i=1}^{k}\left[f(\underline{x}_i)-f(\underline{x}^*)\right] \le \frac{\|\underline{x}_0-\underline{x}^*\|_2^2}{2sk} = O\left(\frac{1}{k}\right). \quad \blacksquare$$

"Classic result: $O(\frac{1}{k})$ of GD."

<u>Thm 2</u>: If $f \in S^{1,1}_{\mu,L}$ and $s \le \frac{2}{L+\mu}$, then

$$\|\underline{x}_k - \underline{x}^*\| \le \left(\sqrt{1 - \frac{2\mu L}{L+\mu}}\right)^k \|\underline{x}_0 - \underline{x}^*\|.$$

Proof: Consider: $\|\underline{x}_{k+1} - \underline{x}^*\|^2$:

$$\|\underline{x}_{k+1} - \underline{x}^*\|^2 = \|\underline{x}_k - \underline{x}^* - s\nabla f(\underline{x}_k)\|^2$$

$$= \|\underline{x}_k - \underline{x}^*\|^2 + s^2\|\nabla f(\underline{x}_k)\|^2 - 2s\nabla f(\underline{x}_k)^T(\underline{x}_k - \underline{x}^*). \qquad (\Delta).$$

<u>Lemma</u> ([Nesterov. Thm 2.1.2]): If $f \in S^{1,1}_{\mu,L}$, then

$$(\nabla f(\underline{x}) - \nabla f(\underline{y}))^T(\underline{x} - \underline{y}) \ge \frac{\mu L}{\mu+L}\|\underline{x} - \underline{y}\|^2 + \frac{1}{\mu+L}\|\nabla f(\underline{x}) - \nabla f(\underline{y})\|^2, \forall \underline{x}, \underline{y} \in \mathbb{R}^n$$

$\nwarrow$ (★)

Using (★), with $\underline{x} = \underline{x}_k$ and $\underline{y} = \underline{x}^*$, noting $\nabla f(\underline{x}^*) = \underline{0}$, we have:

$$\nabla f(\underline{x}_k)^T(\underline{x}_k - \underline{x}^*) \ge \frac{1}{\mu+L}\|\nabla f(\underline{x}_k)\|^2 + \frac{\mu L}{\mu+L}\|\underline{x}_k - \underline{x}^*\|^2.$$

Then: $$(\Delta) \le \left(1 - \frac{2\mu L}{\mu+L}\right)\|\underline{x}_k - \underline{x}^*\|^2 + \underbrace{s\underbrace{\left(s - \frac{2}{\mu+L}\right)}_{\le 0}\|\nabla f(\underline{x}_k)\|^2}_{\le 0}$$

$$\Rightarrow \|\underline{x}_{k+1} - \underline{x}^*\|^2 \le \left(1 - \frac{2\mu L}{\mu+L}\right)\|\underline{x}_k - \underline{x}^*\|^2.$$

Taking square root on both sides. done!

Now, it remains to show (★) is true.

Since $\underset{\subset \mathcal{F}^{1,1}_L}{f \in S^{1,1}_{\mu,L}}$, we have $\forall \underline{x}, \underline{y} \in \mathbb{R}^n$:

1°. $f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2$. (from last lecture)

2° $f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|y-x\|^2$. ($\mu$-strongly convex).

From 1°: interchange x & y: $f(x) \le f(y) + \nabla f(y)^T (x-y) + \frac{L}{2}\|y-x\|^2$.

Adding two copies: $\underline{(\nabla f(x) - \nabla f(y))^T (x-y) \le L\|y-x\|^2}$

By the same token on 2°, we can show,

$$\underline{(\nabla f(x) - \nabla f(y))^T (x-y) \ge \mu\|x-y\|^2}.$$

Now, let's pick $x_0 \in \mathbb{R}^n$ and consider the auxiliary fn:

$$\phi(y) \triangleq f(y) - \nabla f(x_0)^T y. \Rightarrow \nabla\phi(y) = \nabla f(y) - \nabla f(x_0)$$

It's clear that $\phi(\cdot) \in \mathcal{F}_L^{1,1}$, and its opt. pt. is $y^* = x_0$.

(take the grad of $\phi(\cdot)$ & set it to $0 \Rightarrow \nabla\phi(y) = \nabla f(y) - \nabla f(x_0) = 0$

$\Rightarrow y^* = x_0$).

Thus, $\phi(x_0) = \phi(y^*) \le \phi(y - \frac{1}{L}\nabla\phi(y))$.

From 1°: $\underbrace{\phi(y - \frac{1}{L}\nabla\phi(y))}_{\ge \phi(x_0)} - \phi(y) - \underbrace{\nabla\phi(y)^T(\cancel{y} - \frac{1}{L}\phi(y) \cancel{-y})}_{-\frac{1}{L}\|\nabla\phi(y)\|^2} \le \underbrace{\frac{L}{2}\|\cancel{y} - \frac{1}{L}\nabla\phi(y) \cancel{-y}\|^2}_{\frac{1}{2L}\|\nabla\phi(y)\|^2}$

$\Rightarrow \underbrace{\phi(x_0)}_{= f(x_0) - \nabla f(x_0)^T x_0} + \frac{1}{2L}\|\nabla f(y) - \nabla f(x_0)\|^2 \le \phi(y) = f(y) - \nabla f(x_0)^T y$

$\Rightarrow \underline{f(x_0) + \nabla f(x_0)^T (y - x_0) + \frac{1}{2L}\|\nabla f(y) - \nabla f(x_0)\|^2 \le f(y)}.$

Interchange $\underline{x}_0$ & y:

$$f(y)+\nabla f(y)^T(\underline{x}_0-y)+\frac{1}{2L}\|\nabla f(y)-\nabla f(\underline{x}_0)\|^2 \leq f(\underline{x}_0).$$

Adding two copies (rename $\underline{x}_0$ as $\underline{x}$):

$$\Rightarrow \underbrace{\frac{1}{L}\|\nabla f(y)-\nabla f(\underline{x})\|^2 \leq (\nabla f(\underline{x})-\nabla f(y))^T(\underline{x}-y)}. \quad \left(\text{coercity of gradients}\right)$$

In summary: we have

(1) $\frac{1}{L}\|\nabla f(\underline{x})-\nabla f(y)\|^2 \leq (\nabla f(\underline{x})-\nabla f(y))^T(\underline{x}-y)$

(2) $\mu\|\underline{x}-y\|^2 \leq (\nabla f(\underline{x})-\nabla f(y))^T(\underline{x}-y).$

Thus, (1) $\times \frac{L}{\mu+L}$ + (2) $\times \frac{L}{\mu+L}$ $\Rightarrow$

$$\frac{\mu L}{\mu+L}\|\underline{x}-y\|^2+\frac{1}{\mu+L}\|\nabla f(\underline{x})-\nabla f(y)\|^2 \leq \underbrace{\frac{2L}{\mu+L}}(\nabla f(\underline{x})-\nabla f(y))^T(\underline{x}-y).$$

Clearly, if $L=\mu$, then we are done.

So, it remains to show the case w/ $L>\mu$.

Let's consider $d(\underline{x}) \triangleq f(\underline{x})-\frac{1}{2}\mu\|\underline{x}\|^2$. Then, $d(\underline{x}) \in \mathcal{F}^{1,1}_{L-\mu}$.

Note: $\nabla d(\underline{x}) = \nabla f(\underline{x})-\mu\underline{x}$. $\quad$ (☆☆).

Using (1): $\underbrace{(\nabla d(\underline{x})-\nabla d(y))}^T(\underline{x}-y) \geq \frac{1}{L-\mu}\|\nabla d(\underline{x})-\nabla d(y)\|^2 \quad (*)$

Using (☆☆): $\nabla d(\underline{x})-\nabla d(y) = [\nabla f(\underline{x})-\nabla f(y)]-\mu(\underline{x}-y).$

Plugging this into (*) yields:

$$(L-\mu)\left[(\nabla f(x)-\nabla f(y))-\mu(x-y)\right]^T(x-y)$$

$$\geq \|\nabla f(x)-\nabla f(y)\|^2+\mu\|x-y\|^2-2\mu(\nabla f(x)-\nabla f(y))^T(x-y)$$

$$\Rightarrow -\overset{\cancel{\mu^2}\ \ \mu L}{\mu(L-\mu)}\|x-y\|^2+(L-\mu)(\nabla f(x)-\nabla f(y))^T(x-y)$$

$$\geq \|\nabla f(x)-\nabla f(y)\|^2+\cancel{\mu^2\|x-y\|^2}-2\mu(\nabla f(x)-\nabla f(y))^T(x-y)$$

$$\Rightarrow (\overset{1}{\cancel{L}}+\mu)(\nabla f(x)-\nabla f(y))^T(x-y)\geq \underbrace{\|\nabla f(x)-\nabla f(y)\|^2+\mu L\|x-y\|^2}_{L+\mu}$$

Diving $(L+\mu)$ on both sides, we're done! $\blacksquare$

Useful Ineq. in Convex Analysis:

$$\begin{cases} f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|x-y\|^2, & \text{if } f \in \mathcal{F}_L^{1,1}. \\ f(y) \ge f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|x-y\|^2, & \text{if } f \in S_\mu^1 \end{cases}$$

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|, \qquad \text{if } f \in \mathcal{F}_L^{1,1}.$$

$$f(x) + \nabla f(x)^T (y-x) + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \le f(y), \text{ if } f \in \mathcal{F}_L^{1,1}$$

$$f(x) + \nabla f(x)^T (y-x) + \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\|^2 \ge f(y), \text{ if } f \in S_\mu^1$$

Interchanging & Adding:

$$\begin{cases} \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \le (\nabla f(x) - \nabla f(y))^T (x-y) \le L\|x-y\|^2, & \text{if } f \in \mathcal{F}_L^{1,1} \\ \mu\|x-y\|^2 \le (\nabla f(x) - \nabla f(y))^T (x-y) \le \frac{1}{\mu}\|\nabla f(x) - \nabla f(y)\|^2, & \text{if } f \in S_\mu^1. \end{cases}$$

Strongest Result: If $f \in S_{\mu,L}^{1,1}$

$$(\nabla f(x) - \nabla f(y))^T (x-y) \ge \frac{\mu L}{\mu+L}\|x-y\|^2 + \frac{1}{\mu+L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Convex Combinations: $\alpha \in [0,1]$

$$f(\alpha x + (1-\alpha) y) + \frac{\alpha(1-\alpha)}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \le \alpha f(x) + (1-\alpha) f(y)$$

$$\le f(\alpha x + (1-\alpha) y) + \frac{\alpha(1-\alpha)L}{2}\|x-y\|^2.$$