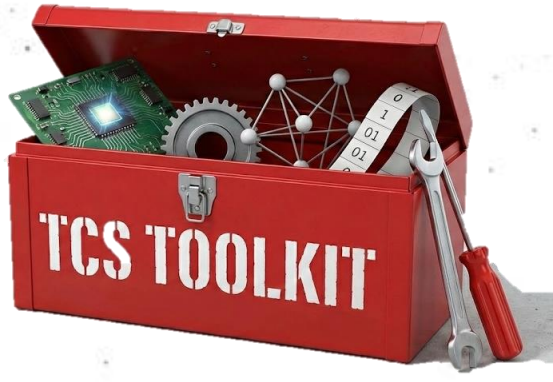
CS 58500 – Theoretical Computer Science Toolkit

Lecture 11 (02/24)

Gradient Descent



https://ruizhezhang.com/course_spring_2026.html

Today’s Lecture

- Lipschitz Optimization
- Smooth Optimization
- Well-Conditioned Optimization

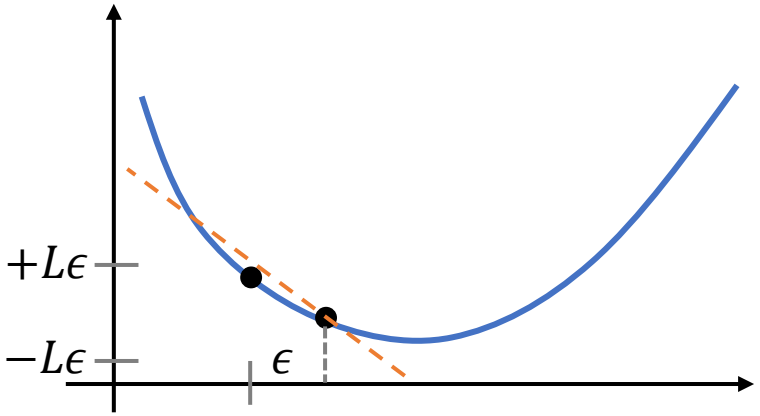
Lipschitz Optimization

- $f: \mathcal{X} \to \mathbb{R}$ is L-Lipschitz if
$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq L \cdot \|\boldsymbol{x} - \boldsymbol{y}\|_2 \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$$
- $\boldsymbol{g}$ is a subgradient of f at $\boldsymbol{x} \in \mathcal{X}$ (denoted as $\boldsymbol{g} \in \partial f(\boldsymbol{x})$) if
$$f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{x}' - \boldsymbol{x} \rangle \quad \forall \boldsymbol{x}' \in \mathcal{X}$$

Lemma (Subgradient bound)**.** Let $f: \mathcal{X} \to \mathbb{R}$ be convex and L-Lipschitz for $\mathcal{X} \subseteq \mathbb{R}^n$. Then, for $\boldsymbol{x} \in \text{relint}(\mathcal{X})$ and $\boldsymbol{g} \in \partial f(\boldsymbol{x})$ contained in the lowest-dimensional subspace containing $\mathcal{X}$, we have $\|\boldsymbol{g}\|_2 \leq L$. Moreover, the converse direction holds as well

Proof.

- Suppose $\exists \boldsymbol{x} \in \mathcal{X}, \boldsymbol{g} \in \partial f(\boldsymbol{x})$ such that $\|\boldsymbol{g}\|_2 > L$
- Consider $\boldsymbol{y} \coloneqq \boldsymbol{x} + \epsilon \boldsymbol{g} \in \mathcal{X}$.
- L-Lipschitz implies that $f(\boldsymbol{y}) \leq f(\boldsymbol{x}) + L\|\boldsymbol{y} - \boldsymbol{x}\|_2 = f(\boldsymbol{x}) + L\epsilon\|\boldsymbol{g}\|_2$



- Subgradient implies that $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{y} - \boldsymbol{x} \rangle = f(\boldsymbol{x}) + \langle \boldsymbol{g}, \epsilon \boldsymbol{g} \rangle > f(\boldsymbol{x}) + L\epsilon\|\boldsymbol{g}\|_2$ (contradiction!)

■

Lipschitz Optimization: Hardness

Theorem 1 (Lipschitz convex lower bound). Let $\epsilon, L, R > 0$. No algorithm $\mathcal{A}$ which accesses a target L-Lipschitz, convex function $f: RB_2^n \to \mathbb{R}$ using a subgradient oracle O and produces iterates $\{\boldsymbol{x}_t\}_{0 \le t < T}$ can optimize f to additive error ϵ using $T < \min\left\{n, \left(\frac{LR}{4\epsilon}\right)^2\right\}$ queries, subject to the restriction that

$$\boldsymbol{x}_0 = \mathbf{0}, \qquad \boldsymbol{x}_t \in \operatorname{span}(\{O(\boldsymbol{x}_s)\}_{0 \le s < t}) \quad \forall t \in [T]$$

Proof.

- Define $f(\boldsymbol{x}) \coloneqq \gamma \max_{i \in [T]} \boldsymbol{x}_i + \frac{\alpha}{2}\|\boldsymbol{x}\|_2^2$ with $\alpha, \gamma > 0$ to be chosen later
- f is convex and $(\gamma + \alpha R)$-Lipschitz $(\gamma + \alpha R \le L)$
- $\partial f(x) = \gamma \cdot \operatorname{conv}\left(\{\boldsymbol{e}_i : i \in \arg\max_{j \in [T]} \boldsymbol{x}_j\}\right) + \alpha \boldsymbol{x}$
- Let $O(\boldsymbol{x}) = \gamma \boldsymbol{e}_i + \alpha \boldsymbol{x}$ with the smallest $i \in \arg\max_{j \in [T]} \boldsymbol{x}_j$

The subgradient of a max function at $\boldsymbol{x}$ is the convex hull of the gradients of the "active" functions at $\boldsymbol{x}$

Lipschitz Optimization: Hardness

Theorem 1 (Lipschitz convex lower bound). Let $\epsilon, L, R > 0$. No algorithm $\mathcal{A}$ which accesses a target L-Lipschitz, convex function $f: RB_2^n \to \mathbb{R}$ using a subgradient oracle O and produces iterates $\{\boldsymbol{x}_t\}_{0 \le t < T}$ can optimize f to additive error ϵ using $T < \min\left\{n, \left(\frac{LR}{4\epsilon}\right)^2\right\}$ queries, subject to the restriction that

$$\boldsymbol{x}_0 = \mathbf{0}, \qquad \boldsymbol{x}_t \in \operatorname{span}(\{O(\boldsymbol{x}_s)\}_{0 \le s < t}) \quad \forall t \in [T]$$

Proof.

- Define $f(\boldsymbol{x}) \coloneqq \gamma \max_{i \in [T]} \boldsymbol{x}_i + \frac{\alpha}{2} \|\boldsymbol{x}\|_2^2$ with $\alpha, \gamma > 0$ to be chosen later
- Let $O(\boldsymbol{x}) = \gamma \boldsymbol{e}_i + \alpha \boldsymbol{x}$ with the smallest $i \in \arg\max_{j \in [T]} \boldsymbol{x}_j$
- $O(\boldsymbol{x}_0) = (\gamma, 0, \dots, 0)$, $\boldsymbol{x}_1 = (?, 0, \dots, 0)$, $O(\boldsymbol{x}_1) = (?, ?, 0, \dots, 0)$
- For all $0 \le t < T$, $\boldsymbol{x}_t$ is only supported in the first t coordinate and $f(\boldsymbol{x}_t) \ge 0$

Lipschitz Optimization: Hardness

Theorem 1 (Lipschitz convex lower bound). Let $\epsilon, L, R > 0$. No algorithm $\mathcal{A}$ which accesses a target L-Lipschitz, convex function $f: RB_2^n \to \mathbb{R}$ using a subgradient oracle O and produces iterates $\{\boldsymbol{x}_t\}_{0 \le t < T}$ can optimize f to additive error ϵ using $T < \min\left\{n, \left(\frac{LR}{4\epsilon}\right)^2\right\}$ queries, subject to the restriction that

$$\boldsymbol{x}_0 = \mathbf{0}, \qquad \boldsymbol{x}_t \in \operatorname{span}(\{O(\boldsymbol{x}_s)\}_{0 \le s < t}) \quad \forall t \in [T]$$

Proof.

- Define $f(\boldsymbol{x}) \coloneqq \gamma \max_{i \in [T]} \boldsymbol{x}_i + \frac{\alpha}{2}\|\boldsymbol{x}\|_2^2$ with $\alpha, \gamma > 0$ to be chosen later
- $\boldsymbol{x}^\star \coloneqq \sum_{i \in [T]} -\frac{\gamma}{\alpha T}\boldsymbol{e}_i$
- $f(\boldsymbol{x}^\star) = -\frac{\gamma^2}{2\alpha T}$ and $\partial f(x^\star) = \sum_{i \in [T]} -\frac{\gamma}{T}\boldsymbol{e}_i + \gamma \cdot \operatorname{conv}(\{\boldsymbol{e}_i : i \in [T]\}) \ni \mathbf{0}$
- Thus, $\boldsymbol{x}^\star$ is a minimizer if $\|\boldsymbol{x}^\star\|_2 \le R \iff \gamma^2 \le R^2\alpha^2 T$

Lipschitz Optimization: Hardness

Theorem 1 (Lipschitz convex lower bound). Let $\epsilon, L, R > 0$. No algorithm $\mathcal{A}$ which accesses a target L-Lipschitz, convex function $f: RB_2^n \to \mathbb{R}$ using a subgradient oracle O and produces iterates $\{\boldsymbol{x}_t\}_{0 \le t < T}$ can optimize f to additive error ϵ using $T < \min\left\{n, \left(\frac{LR}{4\epsilon}\right)^2\right\}$ queries, subject to the restriction that

$$\boldsymbol{x}_0 = \mathbf{0}, \qquad \boldsymbol{x}_t \in \operatorname{span}(\{O(\boldsymbol{x}_s)\}_{0 \le s < t}) \quad \forall t \in [T]$$

Proof.

- Let $\gamma := \frac{L}{2}$ and $\alpha := \frac{L}{2R\sqrt{T}}$, which satisfy the conditions
- $f(\boldsymbol{x}^\star) = -\frac{\gamma^2}{2\alpha T} = -\frac{LR}{4\sqrt{T}}$, while for all $0 \le t < T$, $f(\boldsymbol{x}_t) \ge 0$. That is, the error of $\mathcal{A}$ is $\ge \frac{LR}{4\sqrt{T}}$
- For ϵ-error, we need $\frac{LR}{4\sqrt{T}} < \epsilon$, which gives $T > \left(\frac{LR}{4\epsilon}\right)^2$

■

Lipschitz Optimization: Hardness

Theorem 1 (Lipschitz convex lower bound). Let $\epsilon, L, R > 0$. No algorithm $\mathcal{A}$ which accesses a target L-Lipschitz, convex function $f: RB_2^n \to \mathbb{R}$ using a subgradient oracle O and produces iterates $\{\boldsymbol{x}_t\}_{0 \le t < T}$ can optimize f to additive error ϵ using $T < \min\left\{n, \left(\frac{LR}{4\epsilon}\right)^2\right\}$ queries, subject to the restriction that

$$\boldsymbol{x}_0 = \mathbf{0}, \qquad \boldsymbol{x}_t \in \operatorname{span}(\{O(\boldsymbol{x}_s)\}_{0 \le s < t}) \quad \forall t \in [T]$$

- The cutting-plane method has query complexity $n \log(n/\epsilon)$, matching the first part of the lower bound
- Projected gradient descent can match the second part of the lower bound

Lipschitz Optimization: Projected gradient descent

Update rule of PGD:

$$\boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t), \qquad \boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}(\boldsymbol{x}_t - \eta \boldsymbol{g}_t) \quad \forall 0 \le t < T$$

Here, $\eta > 0$ is the step size, and $\Pi_{\mathcal{X}}$ is the Euclidean projection to $\mathcal{X}$

Theorem (PGD). Let $f: \mathcal{X} \to \mathbb{R}$ be convex and L-Lipschitz for $\mathcal{X} \subseteq RB_2^n$. Suppose $\boldsymbol{x}_0 \leftarrow \mathbf{0}$, $\boldsymbol{x}_{t+1} \leftarrow \Pi_{\mathcal{X}}(\boldsymbol{x}_t - \eta \boldsymbol{g}_t)$ with $\eta = \frac{R}{L\sqrt{T}}$, and let $\overline{\boldsymbol{x}} \coloneqq \frac{1}{T}\sum_{0 \le t < T} \boldsymbol{x}_t$. Then,

$$f(\overline{\boldsymbol{x}}) - \min_{x \in \mathcal{X}} f(\boldsymbol{x}) \le \frac{LR}{\sqrt{T}}$$

- For ϵ-error, $T > \left(\frac{LR}{\epsilon}\right)^2$, which is optimal by the lower bound

Lipschitz Optimization: Projected gradient descent

Theorem (PGD). Let $f: \mathcal{X} \to \mathbb{R}$ be convex and L-Lipschitz for $\mathcal{X} \subseteq RB_2^n$. Suppose $\boldsymbol{x}_0 \leftarrow \mathbf{0}$, $\boldsymbol{x}_{t+1} \leftarrow \Pi_{\mathcal{X}}(\boldsymbol{x}_t - \eta \boldsymbol{g}_t)$ with $\eta = \frac{R}{L\sqrt{T}}$, and let $\overline{\boldsymbol{x}} \coloneqq \frac{1}{T}\sum_{0 \le t < T} \boldsymbol{x}_t$. Then,

$$f(\overline{\boldsymbol{x}}) - \min_{x \in \mathcal{X}} f(\boldsymbol{x}) \le \frac{LR}{\sqrt{T}}$$

Proof.

- Our goal is to upper bound $f(\overline{\boldsymbol{x}}) - f(\boldsymbol{x}^\star)$. By convexity, it suffices to upper bound

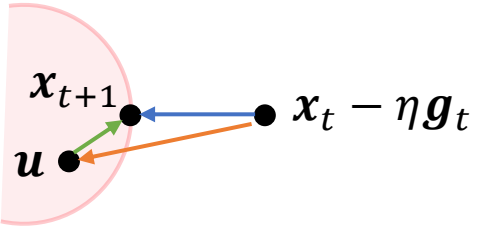
$$f(\overline{\boldsymbol{x}}) - f(\boldsymbol{x}^\star) \le \frac{1}{T}\sum_{0 \le t < T} \big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\big) \le \frac{1}{T}\sum_{0 \le t < T} \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^\star \rangle$$ "potential"

- According to the update rule,

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x} \in \mathcal{X}} \frac{1}{2}\|\boldsymbol{x} - (\boldsymbol{x}_t - \eta \boldsymbol{g}_t)\|_2^2 = \arg\min_{\boldsymbol{x} \in \mathcal{X}} \langle \eta \boldsymbol{g}_t, \boldsymbol{x} \rangle + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2$$

- The first-order optimality condition implies that

$$\langle \eta \boldsymbol{g}_t + (\boldsymbol{x}_{t+1} - \boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{u} \rangle \le 0 \quad \forall \boldsymbol{u} \in \mathcal{X}$$



Lipschitz Optimization: Projected gradient descent

$$\langle \eta \boldsymbol{g}_t + (\boldsymbol{x}_{t+1} - \boldsymbol{x}_t), \boldsymbol{x}_{t+1} - \boldsymbol{u} \rangle \leq 0 \quad \forall \boldsymbol{u} \in \mathcal{X}$$

- We have
$$\begin{aligned}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^\star \rangle &= \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle + \langle \boldsymbol{g}_t, \boldsymbol{x}_{t+1} - \boldsymbol{x}^\star \rangle \\ &\leq \langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}_{t+1} \rangle + \eta^{-1} \langle \boldsymbol{x}_t - \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x}^\star \rangle \\ &\leq \|\boldsymbol{g}_t\|_2 \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|_2 + \eta^{-1} \langle \boldsymbol{x}_t - \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{x}^\star \rangle\end{aligned}$$

- You can check that the following identity holds for any $\boldsymbol{x}_t, \boldsymbol{x}_{t+1}, \boldsymbol{u}$:
$$\langle \boldsymbol{x}_t - \boldsymbol{x}_{t+1}, \boldsymbol{x}_{t+1} - \boldsymbol{u} \rangle = \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{u}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{u}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|_2^2$$

- Thus, we have
$$\begin{aligned}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^\star \rangle &\leq \|\boldsymbol{g}_t\|_2 \|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|_2 - \frac{1}{2\eta}\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|_2^2 + \frac{1}{2\eta}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_2^2 - \frac{1}{2\eta}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|_2^2 \\ &\leq \frac{\eta}{2}\|\boldsymbol{g}_t\|_2^2 + \frac{1}{2\eta}\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_2^2 - \frac{1}{2\eta}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|_2^2 \\ &\leq \frac{\eta L^2}{2} + \frac{1}{2\eta}\left(\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|_2^2\right)\end{aligned}$$

Lipschitz Optimization: Projected gradient descent

$$\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^\star \rangle \leq \frac{\eta L^2}{2} + \frac{1}{2\eta}(\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|_2^2)$$

- Now, we can sum for $0 \leq t < T$:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\langle \boldsymbol{g}_t, \boldsymbol{x}_t - \boldsymbol{x}^\star \rangle &\leq \frac{\eta L^2}{2} + \frac{1}{2\eta T}\sum_{t=0}^{T-1}(\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^\star\|_2^2) \\ &= \frac{\eta L^2}{2} + \frac{1}{2\eta T}(\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2 - \|\boldsymbol{x}_T - \boldsymbol{x}^\star\|_2^2) \\ &\leq \frac{\eta L^2}{2} + \frac{1}{2\eta T}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2 \leq \frac{\eta L^2}{2} + \frac{R^2}{2\eta T}\end{aligned}$$

- If we take $\eta = \frac{R}{L\sqrt{T}}$, we obtain that $\text{LHS} \leq LR/\sqrt{T}$
- Thus, $f(\overline{\boldsymbol{x}}) - \min_{x \in \mathcal{X}} f(\boldsymbol{x}) \leq LR/\sqrt{T}$

Lipschitz Optimization

Let the target function $f: RB_2^n \to \mathbb{R}$ be convex and L-Lipschitz

- Lower bound: $\min\left\{n, \Omega\left(\frac{L^2R^2}{\epsilon^2}\right)\right\}$
- Upper bound:
  - CPM: $\sim n\log(n/\epsilon)$ (high-accuracy)
  - PGD: $\sim L^2R^2/\epsilon^2$ (low-accuracy)

- The lower bound relies on the assumption that the algorithm can only query points within the span of subgradients returned by the oracle
- This assumption can be removed and obtain an information-theoretic lower bound against any randomized algorithm (Agarwal-Bartlett-Ravikumar-Wainwright '12)

Today's Lecture

- Lipschitz Optimization
- Smooth Optimization
- Well-Conditioned Optimization

Smooth Optimization

- Consider an infinitesimal version of the update rule:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t = \boldsymbol{v}_t$$

- By chain rule, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}_t) = \langle \nabla f(\boldsymbol{x}_t), v_t \rangle = -\|\nabla f(\boldsymbol{x}_t)\|_2^2 \quad \text{if } \boldsymbol{v}_t = -\nabla f(\boldsymbol{x}_t)$$

- It indicate that $-\nabla f(\boldsymbol{x}_t)$ is the direction of steepest descent
- We obtain the ODE (so-called the gradient flow):

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t = -\nabla f(x_t)$$

- However, we cannot implement this continuous-time dynamics

Smooth Optimization

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t = -\nabla f(x_t)$$

- For time-discretization, one simple method is the forward Euler method:

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t = -\nabla f\left(\boldsymbol{x}_{t_0}\right) \quad \forall t \in [t_0, t_0 + \eta] \quad \Longrightarrow \quad x_{t_0+\eta} = x_{t_0} - \eta \nabla f\left(x_{t_0}\right)$$

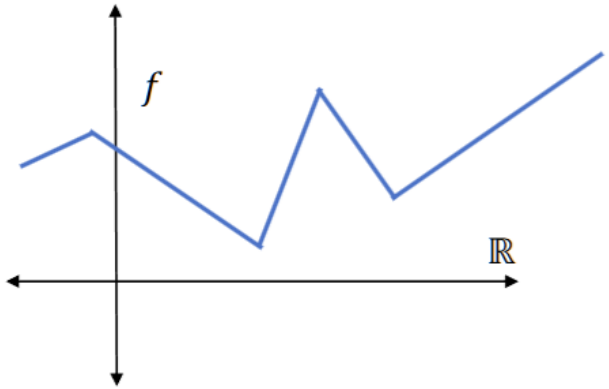
- To guarantee that the discretized dynamics (gradient descent) is close to the continuous time dynamics (gradient flow), we need $\nabla f(\boldsymbol{x})$ to be stable

Smooth Optimization

$f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth, if f is differentiable, and ∇f is L-Lipschitz

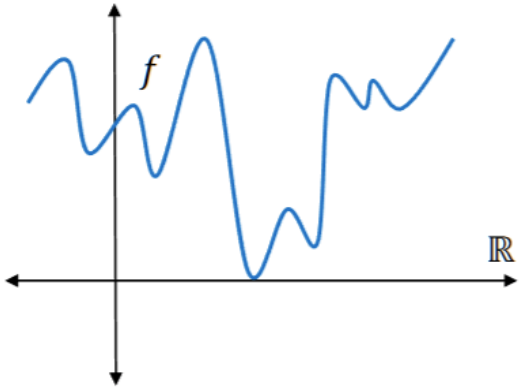
$$|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})| \leq L \cdot \|\boldsymbol{x} - \boldsymbol{y}\| \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

f is L-Lipschitz



(bounded slope)
(bounded 1st derivatives)

f is L-smooth



(bounded curvature)
(bounded 2nd derivative)

(Source: Aaron Sidford)

Smooth Optimization

$f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth, if f is differentiable, and ∇f is L-Lipschitz

$$|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})| \leq L \cdot \|\boldsymbol{x} - \boldsymbol{y}\| \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$$

Lemma. If $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable and convex, then f is L-smooth if and only if

$$f(\boldsymbol{x}') \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2 \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$$

If f is twice-differentiable (and possibly nonconvex), f is L-smooth if and only if

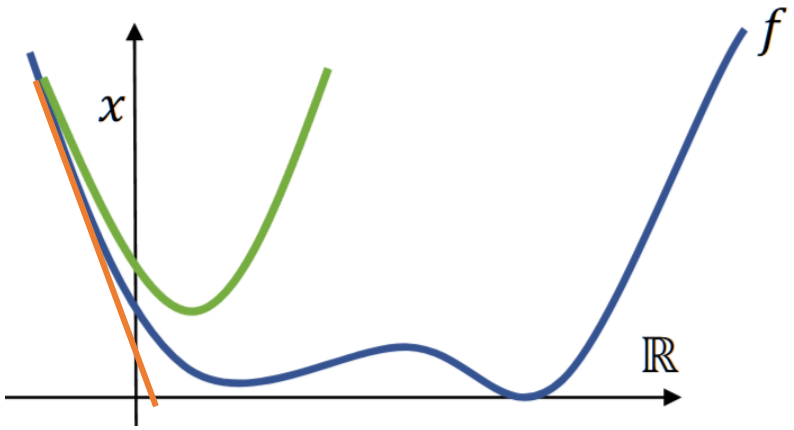
$$|\nabla^2 f(\boldsymbol{x})[\boldsymbol{v}, \boldsymbol{v}]| \coloneqq |\boldsymbol{v}^\top \nabla^2 f(\boldsymbol{x}) \boldsymbol{v}| \leq L \|\boldsymbol{v}\|_2^2 \qquad \forall \boldsymbol{v} \in \mathbb{R}^n$$

Convexity $\Longrightarrow$ linear lower bound

$$f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x} \rangle$$

Smoothness $\Longrightarrow$ quadratic upper bound

$$f(\boldsymbol{x}') \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x} \rangle + \frac{L}{2} \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2$$



Smooth Optimization

Lemma. Let $f: \mathbb{R}^n \to \mathbb{R}$ be L-smooth. Then for any $\boldsymbol{x} \in \mathbb{R}^n$, letting $\boldsymbol{x}' \leftarrow \boldsymbol{x} - \frac{1}{L}\nabla f(\boldsymbol{x})$,

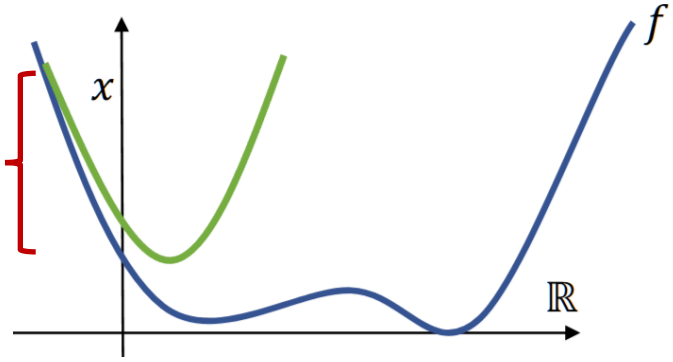
$$f(\boldsymbol{x}') - f(\boldsymbol{x}) \le -\frac{1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2$$

Proof.

- L-smoothness implies an upper bound:

  $$f(\boldsymbol{x}') \le U_{\boldsymbol{x}}(\boldsymbol{x}') \coloneqq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x} \rangle + \frac{L}{2}\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2$$

- Thus, we have $f(\boldsymbol{x}') - f(\boldsymbol{x}) \le U_{\boldsymbol{x}}(\boldsymbol{x}') - U_{\boldsymbol{x}}(\boldsymbol{x}) = \min_{\boldsymbol{x}'} U_{\boldsymbol{x}}(\boldsymbol{x}') - U_{\boldsymbol{x}}(\boldsymbol{x})$



- The first-order optimality condition for RHS is

  $$\nabla f(\boldsymbol{x}) + L(\boldsymbol{x}' - \boldsymbol{x}) = 0 \iff \boldsymbol{x}' = \boldsymbol{x} - L^{-1}\nabla f(\boldsymbol{x})$$

- Then, we have

  $$f(\boldsymbol{x}') \le f(\boldsymbol{x}) - \frac{1}{L}\|\nabla f(\boldsymbol{x})\|_2^2 + \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2 = f(\boldsymbol{x}) - \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2$$

■

Smooth Optimization: Finding Stationary Point

Theorem (Finding stationary point). Let $f: \mathbb{R}^n \to \mathbb{R}$ be L-smooth, let $\epsilon > 0$, and suppose for $\boldsymbol{x}_0 \in \mathbb{R}^n$ we have $f(\boldsymbol{x}_0) - \min f(\boldsymbol{x}) \leq \Delta$. Then iterating $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \frac{1}{L}\nabla f(\boldsymbol{x}_t)$ for $0 \leq t < T$ where $T \geq \frac{2L\Delta}{\epsilon^2}$,

$$\min_{0 \leq t < T} \|\nabla f(\boldsymbol{x}_t)\|_2 \leq \epsilon$$

Proof.

- Suppose the conclusion is not true, i.e., $\|\nabla f(\boldsymbol{x}_t)\|_2 > \epsilon \; \forall t$. Then, by the previous lemma,

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) \leq -\frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|_2^2$$

- Summing over $0 \leq t < T$, we have

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}_0) \leq -\frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|_2^2 < -\frac{T\epsilon^2}{2L} \leq -\Delta$$

■

Smooth Optimization: Finding Stationary Point

Theorem (Finding stationary point). Let $f: \mathbb{R}^n \to \mathbb{R}$ be L-smooth, let $\epsilon > 0$, and suppose for $\boldsymbol{x}_0 \in \mathbb{R}^n$ we have $f(\boldsymbol{x}_0) - \min f(\boldsymbol{x}) \leq \Delta$. Then iterating $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \frac{1}{L}\nabla f(\boldsymbol{x}_t)$ for $0 \leq t < T$ where $T \geq \frac{2L\Delta}{\epsilon^2}$,

$$\min_{0 \leq t < T} \|\nabla f(\boldsymbol{x}_t)\|_2 \leq \epsilon$$

- Finding ϵ-stationary/ϵ-critical point (i.e., $\|\nabla f(\boldsymbol{x})\|_2 \leq \epsilon$) is an important problem in non-convex optimization
- The complexity $\mathcal{O}\left(\frac{L\Delta}{\epsilon^2}\right)$ is optimal (Carmon-Duchi-Hinder-Sidford'20)

Smooth Optimization: Smooth Gradient Descent

Theorem (Smooth GD). Let $f: \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex, and suppose for $\boldsymbol{x}_0 \in \mathbb{R}^n$ we have $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2 \leq R$ for $\boldsymbol{x}^\star \in \arg\min_{x \in \mathbb{R}^n} f(\boldsymbol{x})$. Then iterating $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \frac{1}{L}\nabla f(\boldsymbol{x}_t)$ for $0 \leq t < T$,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \frac{2LR^2}{T}$$

Lemma (Contractivity). Let $f: \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex, and let $\boldsymbol{y} \leftarrow \boldsymbol{x} - \eta\nabla f(\boldsymbol{x})$ for $\eta \leq \frac{1}{L}$. Then for $\boldsymbol{x}^\star \in \arg\min_{x \in \mathbb{R}^n} f(\boldsymbol{x})$, $\|\boldsymbol{y} - \boldsymbol{x}^\star\|_2 \leq \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2$

Proof.

- It suffices to prove that $\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2^2 \geq \|\boldsymbol{x} - \eta\nabla f(\boldsymbol{x}) - \boldsymbol{x}^\star\|_2^2$, which is

$$\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^\star \rangle \geq \frac{\eta}{2}\|\nabla f(\boldsymbol{x})\|_2^2$$

- By convexity, $\langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}^\star \rangle \geq f(\boldsymbol{x}) - f(\boldsymbol{x}^\star) \geq f(\boldsymbol{x}) - f(\boldsymbol{y}) \geq \frac{1}{2L}\|\nabla f(\boldsymbol{x})\|_2^2 \geq \frac{\eta}{2}\|\nabla f(\boldsymbol{x})\|_2^2$

■

Smooth Optimization: Smooth Gradient Descent

Theorem (Smooth GD). Let $f: \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex, and suppose for $\boldsymbol{x}_0 \in \mathbb{R}^n$ we have $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2 \leq R$ for $\boldsymbol{x}^\star \in \arg\min_{x \in \mathbb{R}^n} f(\boldsymbol{x})$. Then iterating $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \frac{1}{L}\nabla f(\boldsymbol{x}_t)$ for $0 \leq t < T$,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \frac{2LR^2}{T}$$

Proof.

- Let $\Phi_t \coloneqq f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)$
- By contractivity, we know that $\|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_2 \leq R \; \forall 0 \leq t \leq T$
- By convexity, $\Phi_t \leq \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \boldsymbol{x}^\star \rangle \leq \|\nabla f(\boldsymbol{x}_t)\|_2 \cdot \|\boldsymbol{x}_t - \boldsymbol{x}^\star\|_2 \leq \|\nabla f(\boldsymbol{x}_t)\|_2 \cdot R$
- On the other hand, we have

$$\Phi_{t+1} - \Phi_t = f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) \leq -\frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|_2^2 \leq -\frac{1}{2LR^2}\Phi_t^2$$

Smooth Optimization: Smooth Gradient Descent

Theorem (Smooth GD). Let $f: \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex, and suppose for $\boldsymbol{x}_0 \in \mathbb{R}^n$ we have $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2 \leq R$ for $\boldsymbol{x}^\star \in \arg\min_{x \in \mathbb{R}^n} f(\boldsymbol{x})$. Then iterating $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \frac{1}{L}\nabla f(\boldsymbol{x}_t)$ for $0 \leq t < T$,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \frac{2LR^2}{T}$$

Proof.

- We have the recursion:

$$\Phi_{t+1} - \Phi_t \leq -\frac{1}{2LR^2}\Phi_t^2$$

- Note that $\frac{1}{\Phi_{t+1}} - \frac{1}{\Phi_t} = \frac{\Phi_t - \Phi_{t+1}}{\Phi_t \Phi_{t+1}} \geq \frac{\Phi_t - \Phi_{t+1}}{\Phi_t^2} \geq \frac{1}{2LR^2}$
- Telescoping for T iterations:

$$\frac{1}{\Phi_T} \geq \frac{1}{\Phi_0} + \frac{T}{2LR^2} \geq \frac{T}{2LR^2}$$

■

Smooth Optimization: Smooth Gradient Descent

Theorem (Smooth GD). Let $f: \mathbb{R}^n \to \mathbb{R}$ be L-smooth and convex, and suppose for $\boldsymbol{x}_0 \in \mathbb{R}^n$ we have $\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2 \leq R$ for $\boldsymbol{x}^\star \in \arg\min_{x \in \mathbb{R}^n} f(\boldsymbol{x})$. Then iterating $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \frac{1}{L}\nabla f(\boldsymbol{x}_t)$ for $0 \leq t < T$,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \frac{2LR^2}{T}$$

- The error bound can be tightened:

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \frac{2LR^2}{T+4}$$

- The gradient descent is provably non-optimal in this setting

Today's Lecture

- Lipschitz Optimization
- Smooth Optimization
- Well-Conditioned Optimization

Well-Conditioned Optimization

- For smooth GD, the convergence rate is $\mathcal{O}(LR^2/\epsilon)$
- Can gradient descent achieve linear convergence rate, i.e., $\sim \log(1/\epsilon)$, or equivalently,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \exp\big(-\Omega(t)\big)$$

- To get some intuition, consider the gradient flow:

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\big) = \frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}_t) = -\|\nabla f(\boldsymbol{x}_t)\|_2^2$$

- Suppose the following condition holds: $\|\nabla f(\boldsymbol{x}_t)\|_2^2 \geq C\big(f(\boldsymbol{x}_t) - f(x^\star)\big)$
- Then, by solving the ODE (or by Grönwall's inequality), we get the linear convergence rate:

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star) \leq e^{-Ct}\big(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)\big)$$

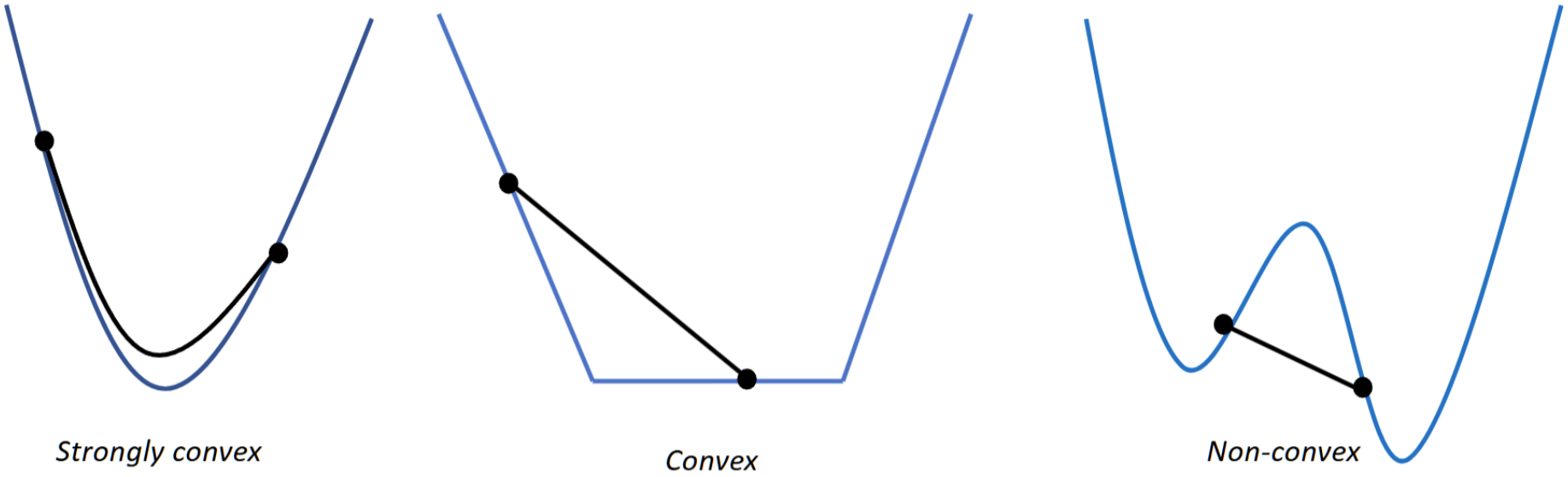
- Another way to see this is to define a Lyapunov function

$$V(t) \coloneqq e^{Ct}\big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\big), \qquad \mathrm{d}V(t) = e^{Ct}\big(C\big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\big) - \|\nabla f(\boldsymbol{x}_t)\|_2^2\big) \leq 0$$

Well-Conditioned Optimization

$f: \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex if

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{\mu\lambda(1-\lambda)}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \qquad \forall \boldsymbol{x} \neq \boldsymbol{y} \in \mathbb{R}^n$$



(Source: Aaron Sidford)

Well-Conditioned Optimization

$f: \mathbb{R}^n \to \mathbb{R}$ is μ-strongly convex if

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}) - \frac{\mu\lambda(1-\lambda)}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2 \qquad \forall \boldsymbol{x} \neq \boldsymbol{y} \in \mathbb{R}^n$$

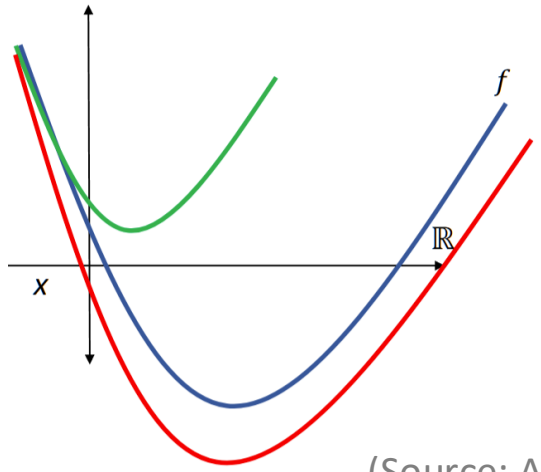
Lemma. If $f: \mathbb{R}^n \to \mathbb{R}$ is differentiable, then f is μ-strongly convex if and only if

$$f(\boldsymbol{x}') \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}' - \boldsymbol{x} \rangle + \frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2 \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$$

If f is twice-differentiable, f is μ-strongly convex iff $\nabla^2 f(\boldsymbol{x})[\boldsymbol{v}, \boldsymbol{v}] \geq \mu\|\boldsymbol{v}\|_2^2 \; \forall \boldsymbol{v} \in \mathbb{R}^n$

Smoothness $\Longrightarrow$ quadratic upper bound

Strong-convexity $\Longrightarrow$ quadratic lower bound



(Source: Aaron Sidford)

Well-Conditioned Optimization

Lemma. Let $f: \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex, and let $\boldsymbol{x}^\star \coloneqq \arg\min f(\boldsymbol{x})$. Then

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^\star) \leq \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|_2^2$$

Proof.

- Strong-convexity implies that for any $\boldsymbol{y} \in \mathbb{R}^n$,

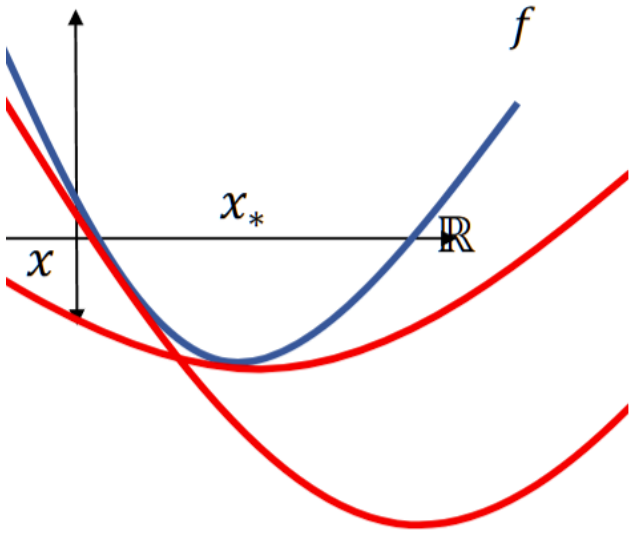
$$f(\boldsymbol{y}) \geq L_{\boldsymbol{x}}(\boldsymbol{y}) \coloneqq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

- Thus, we have

$$f(\boldsymbol{x}^\star) = \min_{\boldsymbol{y}\in\mathbb{R}^n} f(\boldsymbol{y}) \geq \min_{\boldsymbol{y}\in\mathbb{R}^n} L_{\boldsymbol{x}}(\boldsymbol{y})$$

$$= f(\boldsymbol{x}) + \left\langle \nabla f(\boldsymbol{x}), -\frac{1}{\mu}\nabla f(\boldsymbol{x}) \right\rangle + \frac{\mu}{2}\left\|\frac{1}{\mu}\nabla f(\boldsymbol{x})\right\|_2^2$$

$$= f(\boldsymbol{x}) - \frac{1}{2\mu}\|\nabla f(\boldsymbol{x})\|_2^2$$



■

Well-Conditioned Optimization: Well-conditioned Gradient Descent

Theorem (Well-conditioned GD). Let $f: \mathbb{R}^n \to \mathbb{R}$ be L-smooth and μ-strongly convex, and let $x^\star := \arg\min f(x)$ and $\kappa := \frac{L}{\mu} \geq 1$. Then iterating $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \frac{1}{L}\nabla f(\boldsymbol{x}_t)$ for $0 \leq t < T$,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \left(1 - \frac{1}{\kappa}\right)^T \left(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)\right)$$

Proof.

- For $0 \leq t < T$, we have

$$f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) \leq -\frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|_2^2 \leq -\frac{\mu}{L}\left(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\right)$$

L-smooth μ-strongly convex

- Thus,

$$f(\boldsymbol{x}_{t+1}) - f(x^\star) \leq \left(1 - \frac{1}{\kappa}\right)\left(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^\star)\right)$$

■

Well-Conditioned Optimization: Well-conditioned Gradient Descent

Theorem (Well-conditioned GD). Let $f: \mathbb{R}^n \to \mathbb{R}$ be L-smooth and μ-strongly convex, and let $x^\star := \arg\min f(x)$ and $\kappa := \frac{L}{\mu} \geq 1$. Then iterating $\boldsymbol{x}_{t+1} \leftarrow \boldsymbol{x}_t - \frac{1}{L}\nabla f(\boldsymbol{x}_t)$ for $0 \leq t < T$,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \leq \left(1 - \frac{1}{\kappa}\right)^T \left(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)\right)$$

- GD can achieve linear convergence rate $T = \mathcal{O}(\kappa \log(1/\epsilon))$
- This result only uses the smoothness and the bound $C\left(f(\boldsymbol{x}) - f(\boldsymbol{x}^\star)\right) \leq \|\nabla f(\boldsymbol{x})\|_2^2$, which can still hold even if f is non-convex
- Bounds of this form is called Polyak-Łojasiewicz conditions

Well-Conditioned Optimization: Hardness

Theorem. Let $\kappa \geq 1$ and $\epsilon \in (0,1)$. No algorithm $\mathcal{A}$ which accesses a target L-smooth, μ-strongly convex function f with $\kappa = \frac{L}{\mu}$ using a gradient oracle O can optimize f to additive error $\epsilon(f(\boldsymbol{x}_0) - \min f(\boldsymbol{x}))$ using $T < \frac{\sqrt{\kappa}-1}{2}\log\left(\frac{1}{\kappa\epsilon}\right)$ queries, subject to the restriction that

$$\boldsymbol{x}_0 = \mathbf{0}, \qquad \boldsymbol{x}_t \in \operatorname{span}(\{O(\boldsymbol{x}_s)\}_{0 \leq s < t}) \quad \forall t \in [T]$$

Proof.

- Define the hard function

$$f(x) \coloneqq -\frac{L-\mu}{4}x_1 + \frac{L-\mu}{8}\sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + \frac{L-\mu}{8}x_1^2 + \frac{\mu}{2}\sum_{i=1}^{n}x_i^2 + \frac{\sqrt{L\mu}-\mu}{4}x_n^2$$

- Note that $f = (\text{linear}) + \frac{1}{2}\boldsymbol{x}^\top A\boldsymbol{x}$, and thus, $\nabla^2 f(\boldsymbol{x}) = A$

Well-Conditioned Optimization: Hardness

$$\boldsymbol{v}^2\nabla^2 f(\boldsymbol{x})\boldsymbol{v} = \frac{L-\mu}{4}\sum_{i=1}^{n-1}(v_i - v_{i+1})^2 + \frac{L-\mu}{4}v_1^2 + \mu\sum_{i=1}^{n} v_i^2 + \frac{\sqrt{L\mu}-\mu}{2}v_n^2$$

- L-smooth:

$$\boldsymbol{v}^2\nabla^2 f(\boldsymbol{x})\boldsymbol{v} \leq \frac{L-\mu}{4}\sum_{i=1}^{n-1}\left(2v_i^2 + 2v_{i+1}^2\right)^2 + \frac{L-\mu}{4}v_1^2 + \mu\sum_{i=1}^{n} v_i^2 + \frac{\sqrt{L\mu}-\mu}{2}v_n^2$$

$$\leq (L-\mu)\sum_{i=1}^{n} v_i^2 + \mu\sum_{i=1}^{n} v_i^2 = L\sum_{i=1}^{n} v_i^2$$

- μ-strongly convex:

$$\boldsymbol{v}^2\nabla^2 f(\boldsymbol{x})\boldsymbol{v} \geq \mu\sum_{i=1}^{n} v_i^2$$

Well-Conditioned Optimization: Hardness

$$f(x) \coloneqq -\frac{L-\mu}{4}x_1 + \frac{L-\mu}{8}\sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + \frac{L-\mu}{8}x_1^2 + \frac{\mu}{2}\sum_{i=1}^{n}x_i^2 + \frac{\sqrt{L\mu}-\mu}{4}x_n^2$$

- Note that $\nabla f(\boldsymbol{x}_0) = \nabla f(\mathbf{0}) = (?, 0, \dots, 0)$ and $\boldsymbol{x}_1 = (?, 0, \dots, 0)$ by our assumption
- Then, $\nabla f(\boldsymbol{x}_1) = (?, ?, 0, \dots, 0)$
- By induction, you can show that only the first k coordinates of x_k are non-zero
- By strong convexity, we have

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \geq \frac{\mu}{2}\|\boldsymbol{x}_T - \boldsymbol{x}^\star\|_2^2 \geq \frac{\mu}{2}\sum_{i=T+1}^{n}(\boldsymbol{x}_i^\star)^2$$

Well-Conditioned Optimization: Hardness

$$f(x) \coloneqq -\frac{L-\mu}{4}x_1 + \frac{L-\mu}{8}\sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + \frac{L-\mu}{8}x_1^2 + \frac{\mu}{2}\sum_{i=1}^{n}x_i^2 + \frac{\sqrt{L\mu}-\mu}{4}x_n^2$$

- Let $\boldsymbol{x}^\star \coloneqq \arg\min f(\boldsymbol{x})$. By the first-order optimality condition, $\nabla f(\boldsymbol{x}^\star) = \mathbf{0}$, i.e.,

$$-\frac{L-\mu}{4} + \frac{L-\mu}{4}(x_1^\star - x_2^\star) + \frac{L-\mu}{4}x_1^\star + \mu x_1^\star = 0$$

$$\frac{L-\mu}{4}(x_i^\star - x_{i+1}^\star) + \frac{L-\mu}{4}(x_i^\star - x_i^\star) + \mu x_i^\star = 0 \quad \forall i \in \{2, \dots, n-1\}$$

$$\frac{L-\mu}{4}(x_n^\star - x_{n-1}^\star) + \mu x_n^\star + \frac{\sqrt{L\mu}-\mu}{2}x_n^\star = 0$$

- Solving the linear system, we get that

$$x_i^\star = \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^i \qquad \forall i \in [n]$$

Well-Conditioned Optimization: Hardness

$$f(x) \coloneqq -\frac{L-\mu}{4}x_1 + \frac{L-\mu}{8}\sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + \frac{L-\mu}{8}x_1^2 + \frac{\mu}{2}\sum_{i=1}^{n}x_i^2 + \frac{\sqrt{L\mu}-\mu}{4}x_n^2$$

- By induction, you can show that only the first k coordinates of x_k are non-zero
- By strong convexity, we have

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \geq \frac{\mu}{2}\|\boldsymbol{x}_T - \boldsymbol{x}^\star\|_2^2 \geq \frac{\mu}{2}\sum_{i=T+1}^{n}(\boldsymbol{x}_i^\star)^2 = \frac{\mu}{2}\sum_{i=T+1}^{n}\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2i}$$

$$\overset{n \to \infty}{\approx} \frac{\mu}{2}\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2T}\sum_{i=1}^{n}\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2i} = \frac{\mu}{2}\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2$$

$$\geq \frac{1}{\kappa}\left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2T}\big(f(\boldsymbol{x}_0) - f(\boldsymbol{x}^\star)\big)$$

L-smooth

■

Well-Conditioned Optimization: Hardness

- The hardness result indicates that the optimal query complexity in the well-conditioned setting should be $T = \sqrt{\kappa}\log(1/\epsilon)$, while GD can only achieve $\mathcal{O}(\kappa\log(1/\epsilon))$, which is sub-optimal
- To achieve the optimal rate, we need to use some "history information" (i.e., momentum) in the update rule, and the algorithms are called accelerated GD (first discovered by Nesterov)
- The hard function in the proof is

$$f(x) := -\frac{L-\mu}{4}x_1 + \frac{L-\mu}{8}\sum_{i=1}^{n-1}(x_i - x_{i+1})^2 + \frac{L-\mu}{8}x_1^2 + \frac{\mu}{2}\sum_{i=1}^{n}x_i^2 + \frac{\sqrt{L\mu}-\mu}{4}x_n^2$$

It is related to the Laplacian of a path graph

- Given any convex function, we construct the dependence graph G on the set of variables x_i by connecting $x_i \sim x_j$ if $\nabla f(x)_i$ depends on x_j or $\nabla f(x)_j$ depends on x_i
- GD can only transmit information from one vertex to another in each iteration

Reduction between Smooth and Well-Conditioned Optimizations

Lemma. Suppose that algorithm $\mathcal{A}$ can optimize an L-smooth, convex $f: \mathbb{R}^n \to \mathbb{R}$ with $\boldsymbol{x}^\star \in \arg\min f(\boldsymbol{x})$, and outputs a point $\boldsymbol{x}_T \in \mathbb{R}^n$ using T queries such that,

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^\star) \lesssim \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|_2^2}{T^c}$$

Then there is an algorithm $\mathcal{A}'$ that can optimize an L-smooth, μ-strongly convex $g: \mathbb{R}^n \to \mathbb{R}$ with $\boldsymbol{y}^\star = \arg\min g(\boldsymbol{y})$ and $\kappa := \frac{L}{\mu}$, and outputs a point $\boldsymbol{y}_T \in \mathbb{R}^n$ such that

$$g(\boldsymbol{y}_T) - g(\boldsymbol{y}^\star) \le \epsilon\big(g(\boldsymbol{y}_0) - g(\boldsymbol{y}^\star)\big)$$

in $\mathcal{O}\big(\kappa^{1/c}\log(1/\epsilon)\big)$ queries

- The converse direction also holds (Allen-Zhu-Hazan '16)
- GD for smooth function can achieve $c = 1$, and our lower bound for well-conditioned case shows that $c \le 2$ (and $c = 2$ can be achieved by AGD). Thus, GD is also sub-optimal in smooth optimization

Reduction between Smooth and Well-Conditioned Optimizations

Proof.

- Let $T' \coloneqq \mathcal{O}\left((4\kappa)^{1/c}\right)$.
- We first show that by running $\mathcal{A}$ with T' queries, the error can be decreased by a half
- By the convergence rate of $\mathcal{A}$, we have

$$g(\boldsymbol{y}_T) - g(\boldsymbol{y}^\star) \lesssim \frac{L\|\boldsymbol{y}_0 - \boldsymbol{y}^\star\|_2^2}{\left((4\kappa)^{1/c}\right)^c} = \frac{\mu}{4}\|\boldsymbol{y}_0 - \boldsymbol{y}^\star\|_2^2 \leq \frac{1}{2}\big(g(\boldsymbol{y}_0) - g(\boldsymbol{y}^\star)\big)$$

- Then, by iterating $\log(1/\epsilon)$ times, the error can be decreased to ϵ

■

Summary

Regularity	Goal	Algorithm	Iterations
$f: B_2^n \to \mathbb{R}$, L-Lipschitz	ϵ-optimal	Any algorithm	$\Omega\left((L/\epsilon)^n\right)$
$f: RB_2^n \to \mathbb{R}$, convex, L-Lipschitz	ϵ-optimal	Any first-order method	$\Omega(\min\{n, (LR/\epsilon)^2\})$
		Cutting-plane method	$\tilde{\mathcal{O}}(n)$
		Projected GD	$\mathcal{O}\left((LR/\epsilon)^2\right)$
$f: \mathbb{R}^n \to \mathbb{R}$, L-smooth	ϵ-critical	GD	$\mathcal{O}(L(f(\boldsymbol{x}_0) - f^\star)/\epsilon)$
$f: \mathbb{R}^n \to \mathbb{R}$, convex, L-smooth	ϵ-optimal	GD	$\mathcal{O}(L\|\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|\|_2^2/\epsilon)$
		Any first-order method	$\Omega\left(\sqrt{L\|\|\boldsymbol{x}_0 - \boldsymbol{x}^\star\|\|_2^2/\epsilon}\right)$
$f: \mathbb{R}^n \to \mathbb{R}$, L-smooth, μ-strongly convex	ϵ-optimal	GD	$\tilde{\mathcal{O}}(L/\mu)$
		Any first-order method	$\Omega\left(\sqrt{L/\mu}\right)$