

Convex Optimization and Gradient Descent

Dimitris Fotakis

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING
NATIONAL TECHNICAL UNIVERSITY OF ATHENS, GREECE

Convex Sets and Convex Functions

- Set $S \subseteq \mathbb{R}^d$ is **convex**, if $\forall \mathbf{x}, \mathbf{y} \in S$ and $\forall \lambda \in [0, 1]$,

$$\lambda \mathbf{x} + (1 - \lambda) \mathbf{y} \in S$$

Convex Sets and Convex Functions

- Set $S \subseteq \mathbb{R}^d$ is **convex**, if $\forall \mathbf{x}, \mathbf{y} \in S$ and $\forall \lambda \in [0, 1]$,

$$\lambda \mathbf{x} + (1 - \lambda) \mathbf{y} \in S$$

- Function $f : S \rightarrow \mathbb{R}$ is **convex**, if $\forall \mathbf{x}, \mathbf{y} \in S$ and $\forall \lambda \in [0, 1]$,

$$\lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) \geq f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y})$$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

Convex Sets and Convex Functions

- Set $S \subseteq \mathbb{R}^d$ is **convex**, if $\forall \mathbf{x}, \mathbf{y} \in S$ and $\forall \lambda \in [0, 1]$,

$$\lambda \mathbf{x} + (1 - \lambda) \mathbf{y} \in S$$

- Function $f : S \rightarrow \mathbb{R}$ is **convex**, if $\forall \mathbf{x}, \mathbf{y} \in S$ and $\forall \lambda \in [0, 1]$,

$$\lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}) \geq f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y})$$

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

- Function $f : S \rightarrow \mathbb{R}$ is **α -strongly convex**, if $\forall \mathbf{x}, \mathbf{y} \in S$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{\alpha \|\mathbf{y} - \mathbf{x}\|^2}{2}$$

Local Search - Local Opt is Global Opt

Local Search

Input: convex set S , convex fun. $f : S \rightarrow \mathbb{R}$, initial $x_1 \in S$, radius $\varepsilon > 0$

Neighborhood: $N_\varepsilon(x) = \{y \in S : \|x - y\| \leq \varepsilon\}$

For each $t = 1, 2, \dots$ do:

- If $\exists x \in N_\varepsilon(x_t)$ with $f(x) < f(x_t)$, set $x_{t+1} = x$.
- Else return $x^* = x_t$ as **minimizer** of $f(x)$ in S .

Local Search - Local Opt is Global Opt

Local Search

Input: convex set S , convex fun. $f : S \rightarrow \mathbb{R}$, initial $x_1 \in S$, radius $\varepsilon > 0$

Neighborhood: $N_\varepsilon(x) = \{y \in S : \|x - y\| \leq \varepsilon\}$

For each $t = 1, 2, \dots$ do:

- If $\exists x \in N_\varepsilon(x_t)$ with $f(x) < f(x_t)$, set $x_{t+1} = x$.
- Else return $x^* = x_t$ as **minimizer** of $f(x)$ in S .

Local Optimum is Global Optimum, under Convexity

If set S is convex and f is convex on S , local optimum x^* is **minimizer**:

$$f(x^*) \leq f(z), \text{ for all } z \in S.$$

Local Search - Local Opt is Global Opt

Local Optimum is Global Optimum, under Convexity

If set S is convex and f is convex on S , x^* of local search is **minimizer**:

$$f(x^*) \leq f(z), \text{ for all } z \in S.$$

Proof (by contradiction):

Local Search - Local Opt is Global Opt

Local Optimum is Global Optimum, under Convexity

If set S is convex and f is convex on S , \mathbf{x}^* of local search is **minimizer**:

$$f(\mathbf{x}^*) \leq f(\mathbf{z}), \text{ for all } \mathbf{z} \in S.$$

Proof (by contradiction):

- Let $\mathbf{z}^* \in S$ with $f(\mathbf{z}^*) < f(\mathbf{x}^*)$. Wlog. $\mathbf{z}^* \notin N_\varepsilon(\mathbf{x}^*)$.

Local Search - Local Opt is Global Opt

Local Optimum is Global Optimum, under Convexity

If set S is convex and f is convex on S , \mathbf{x}^* of local search is **minimizer**:

$$f(\mathbf{x}^*) \leq f(\mathbf{z}), \text{ for all } \mathbf{z} \in S.$$

Proof (by contradiction):

- Let $\mathbf{z}^* \in S$ with $f(\mathbf{z}^*) < f(\mathbf{x}^*)$. Wlog. $\mathbf{z}^* \notin N_\varepsilon(\mathbf{x}^*)$.
- Let $\mathbf{y} \in (\lambda\mathbf{x}^* + (1 - \lambda)\mathbf{z}^*)_{\lambda \in [0,1]}$ with $\|\mathbf{y} - \mathbf{x}^*\| = \varepsilon$.

Local Search - Local Opt is Global Opt

Local Optimum is Global Optimum, under Convexity

If set S is convex and f is convex on S , \mathbf{x}^* of local search is **minimizer**:

$$f(\mathbf{x}^*) \leq f(\mathbf{z}), \text{ for all } \mathbf{z} \in S.$$

Proof (by contradiction):

- Let $\mathbf{z}^* \in S$ with $f(\mathbf{z}^*) < f(\mathbf{x}^*)$. Wlog. $\mathbf{z}^* \notin N_\varepsilon(\mathbf{x}^*)$.
- Let $\mathbf{y} \in (\lambda\mathbf{x}^* + (1 - \lambda)\mathbf{z}^*)_{\lambda \in [0,1]}$ with $\|\mathbf{y} - \mathbf{x}^*\| = \varepsilon$.
- Since $\varepsilon > 0$, $\mathbf{y} = \lambda_\varepsilon\mathbf{x}^* + (1 - \lambda_\varepsilon)\mathbf{z}^*$, for some $\lambda_\varepsilon > 0$.

Local Search - Local Opt is Global Opt

Local Optimum is Global Optimum, under Convexity

If set S is convex and f is convex on S , \mathbf{x}^* of local search is **minimizer**:

$$f(\mathbf{x}^*) \leq f(\mathbf{z}), \text{ for all } \mathbf{z} \in S.$$

Proof (by contradiction):

- Let $\mathbf{z}^* \in S$ with $f(\mathbf{z}^*) < f(\mathbf{x}^*)$. Wlog. $\mathbf{z}^* \notin N_\varepsilon(\mathbf{x}^*)$.
- Let $\mathbf{y} \in (\lambda\mathbf{x}^* + (1 - \lambda)\mathbf{z}^*)_{\lambda \in [0,1]}$ with $\|\mathbf{y} - \mathbf{x}^*\| = \varepsilon$.
- Since $\varepsilon > 0$, $\mathbf{y} = \lambda_\varepsilon\mathbf{x}^* + (1 - \lambda_\varepsilon)\mathbf{z}^*$, for some $\lambda_\varepsilon > 0$.
- By convexity of S , $\mathbf{y} \in S$ and thus, $\mathbf{y} \in N_\varepsilon(\mathbf{x}^*)$.

Local Search - Local Opt is Global Opt

Local Optimum is Global Optimum, under Convexity

If set S is convex and f is convex on S , \mathbf{x}^* of local search is **minimizer**:

$$f(\mathbf{x}^*) \leq f(\mathbf{z}), \text{ for all } \mathbf{z} \in S.$$

Proof (by contradiction):

- Let $\mathbf{z}^* \in S$ with $f(\mathbf{z}^*) < f(\mathbf{x}^*)$. Wlog. $\mathbf{z}^* \notin N_\varepsilon(\mathbf{x}^*)$.
- Let $\mathbf{y} \in (\lambda\mathbf{x}^* + (1 - \lambda)\mathbf{z}^*)_{\lambda \in [0,1]}$ with $\|\mathbf{y} - \mathbf{x}^*\| = \varepsilon$.
- Since $\varepsilon > 0$, $\mathbf{y} = \lambda_\varepsilon\mathbf{x}^* + (1 - \lambda_\varepsilon)\mathbf{z}^*$, for some $\lambda_\varepsilon > 0$.
- By convexity of S , $\mathbf{y} \in S$ and thus, $\mathbf{y} \in N_\varepsilon(\mathbf{x}^*)$.
- By convexity of f ,

$$\begin{aligned} f(\mathbf{y}) &= f(\lambda_\varepsilon\mathbf{x}^* + (1 - \lambda_\varepsilon)\mathbf{z}^*) \\ &\leq \lambda_\varepsilon f(\mathbf{x}^*) + (1 - \lambda_\varepsilon)f(\mathbf{z}^*) && \text{convexity of } f \\ &< \lambda_\varepsilon f(\mathbf{x}^*) + (1 - \lambda_\varepsilon)f(\mathbf{x}^*) = f(\mathbf{x}^*) && \text{contradiction!} \end{aligned}$$

Gradient Descent

(Projected) Gradient Descent

Input: convex set S , convex fun. $f : S \rightarrow \mathbb{R}$, $\mathbf{x}_1 \in S$, step size $\eta > 0$

For each $t = 1, 2, \dots, T$ do:

- Update $\mathbf{y}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$
- Project $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{y}_{t+1}\|$

Gradient Descent

(Projected) Gradient Descent

Input: convex set S , convex fun. $f : S \rightarrow \mathbb{R}$, $\mathbf{x}_1 \in S$, step size $\eta > 0$

For each $t = 1, 2, \dots, T$ do:

- Update $\mathbf{y}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$
- Project $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{y}_{t+1}\|$

Ignore project step for the analysis: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$

We let $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$, for brevity. Let \mathbf{x}^* be **minimizer** of f in S .

Gradient Descent

(Projected) Gradient Descent

Input: convex set S , convex fun. $f : S \rightarrow \mathbb{R}$, $\mathbf{x}_1 \in S$, step size $\eta > 0$

For each $t = 1, 2, \dots, T$ do:

- Update $\mathbf{y}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$
- Project $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{y}_{t+1}\|$

Ignore project step for the analysis: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$

We let $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$, for brevity. Let \mathbf{x}^* be **minimizer** of f in S .

By convexity of f , $f(\mathbf{x}^*) \geq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)(\mathbf{x}^* - \mathbf{x}_t)$, we get that:

$$\text{Loss}_{GD} = \sum_{t=1}^T (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \sum_{t=1}^T \mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) \quad (1)$$

Gradient Descent

(Projected) Gradient Descent

Input: convex set S , convex fun. $f : S \rightarrow \mathbb{R}$, $\mathbf{x}_1 \in S$, step size $\eta > 0$

For each $t = 1, 2, \dots, T$ do:

- Update $\mathbf{y}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$
- Project $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in S} \|\mathbf{x} - \mathbf{y}_{t+1}\|$

Ignore project step for the analysis: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$

We let $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$, for brevity. Let \mathbf{x}^* be **minimizer** of f in S .

By convexity of f , $f(\mathbf{x}^*) \geq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)(\mathbf{x}^* - \mathbf{x}_t)$, we get that:

$$\text{Loss}_{GD} = \sum_{t=1}^T (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \sum_{t=1}^T \mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) \quad (1)$$

We show that $\text{Loss}_{GD}/T \rightarrow 0$, as $T \rightarrow \infty$: the **average** of $f(\mathbf{x}_1), \dots, f(\mathbf{x}_T)$ (and $f((\mathbf{x}_1 + \dots + \mathbf{x}_T)/T)$) **converges** to optimum $f(\mathbf{x}^*)$.

Gradient Descent

From the update rule $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$, we get that:

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_t - \eta \mathbf{g}_t - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta \mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) + \eta^2 \|\mathbf{g}_t\|^2 \Rightarrow\end{aligned}$$

Gradient Descent

From the update rule $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$, we get that:

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_t - \eta \mathbf{g}_t - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta \mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) + \eta^2 \|\mathbf{g}_t\|^2 \Rightarrow \\ \mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) &= \frac{1}{2\eta} (\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2) + \frac{\eta}{2} \|\mathbf{g}_t\|^2\end{aligned}\quad (2)$$

Gradient Descent

From the update rule $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$, we get that:

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_t - \eta \mathbf{g}_t - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta \mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) + \eta^2 \|\mathbf{g}_t\|^2 \Rightarrow \\ \mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) &= \frac{1}{2\eta} (\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2) + \frac{\eta}{2} \|\mathbf{g}_t\|^2\end{aligned}\quad (2)$$

Substituting (2) in (1) results in a **telescopic** sum. So, we get that:

$$\begin{aligned}\text{Loss}_{GD} &= \sum_{t=1}^T (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|\mathbf{g}_t\|^2 \\ &\leq \frac{B^2}{2\eta} + \frac{\eta TG^2}{2} \stackrel{\eta = \frac{B}{G\sqrt{T}}}{=} BG\sqrt{T}\end{aligned}$$

Similar bound with step $\eta_t = \frac{B}{G\sqrt{t}}$, because $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$.

Gradient Descent

From the update rule $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$, we get that:

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &= \|\mathbf{x}_t - \eta \mathbf{g}_t - \mathbf{x}^*\|^2 \\ &= \|\mathbf{x}_t - \mathbf{x}^*\|^2 - 2\eta \mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) + \eta^2 \|\mathbf{g}_t\|^2 \Rightarrow \\ \mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) &= \frac{1}{2\eta} (\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2) + \frac{\eta}{2} \|\mathbf{g}_t\|^2\end{aligned}\quad (2)$$

Substituting (2) in (1) results in a **telescopic** sum. So, we get that:

$$\begin{aligned}\text{Loss}_{GD} &= \sum_{t=1}^T (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \frac{\|\mathbf{x}_1 - \mathbf{x}^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|\mathbf{g}_t\|^2 \\ &\leq \frac{B^2}{2\eta} + \frac{\eta T G^2}{2} \stackrel{\eta = \frac{B}{G\sqrt{T}}}{=} B G \sqrt{T}\end{aligned}$$

Similar bound with step $\eta_t = \frac{B}{G\sqrt{t}}$, because $\sum_{t=1}^T \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$.

So, $\text{Loss}_{GD} \leq \varepsilon$, for $T = B^2 G^2 / \varepsilon^2$, where $B = \max_{\mathbf{x}, \mathbf{y} \in S} \|\mathbf{x} - \mathbf{y}\|$ is the **diameter** of S and $G \geq \|\mathbf{g}_t\|$ bounds the norm of f 's **gradient**.

α -Strongly Convex Functions

Using α -strong convexity

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)(\mathbf{x}^* - \mathbf{x}_t) + \frac{\alpha}{2} \|\mathbf{x}^* - \mathbf{x}_t\|^2,$$

we get that:

$$2 \sum_{t=1}^T (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \sum_{t=1}^T \left(2\mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) - \alpha \|\mathbf{x}_t - \mathbf{x}^*\|^2 \right) \quad (3)$$

α -Strongly Convex Functions

Using α -strong convexity

$$f(\mathbf{x}^*) \geq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)(\mathbf{x}^* - \mathbf{x}_t) + \frac{\alpha}{2} \|\mathbf{x}^* - \mathbf{x}_t\|^2,$$

we get that:

$$2 \sum_{t=1}^T (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \sum_{t=1}^T \left(2\mathbf{g}_t(\mathbf{x}_t - \mathbf{x}^*) - \alpha \|\mathbf{x}_t - \mathbf{x}^*\|^2 \right) \quad (3)$$

Substituting (2) in (3), we get that:

$$\begin{aligned} 2\text{Loss}_{GD} &= 2 \sum_{t=1}^T (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\ &\leq \sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}^*\|^2 \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - \alpha \right) + \sum_{t=1}^T \eta_t \|\mathbf{g}_t\|^2 \\ &\stackrel{\eta_t = 1/(\alpha t)}{\leq} 0 + \frac{G^2(1 + \ln T)}{\alpha} \end{aligned}$$