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 $\forall \boldsymbol{x}, \boldsymbol{y} \in S$ and $\forall \lambda \in [0, 1]$,

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

Convex Sets and Convex Functions

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

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

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

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

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

Convex Sets and Convex Functions

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

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

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

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

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

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

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

Local Search - Local Opt is Global Opt

Local Search

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

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

For each $t = 1, 2, \ldots$ 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. $f : S \to \mathbb{R}$, initial $x_1 \in S$, radius $\varepsilon > 0$
Neighborhood: $N_\varepsilon(x) = \{y \in S : \|x - y\| \leq \varepsilon\}$

For each $t = 1, 2, \ldots$ 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, x^* of local search is **minimizer**:

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

Proof (by contradiction):

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

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, x^* of local search is **minimizer**:

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

Proof (by contradiction):

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

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, x^* of local search is **minimizer**:

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

Proof (by contradiction):

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

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

Proof (by contradiction):

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

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, x^* of local search is **minimizer**:

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

Proof (by contradiction):

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

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

Gradient Descent

(Projected) Gradient Descent

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

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

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

Gradient Descent

(Projected) Gradient Descent

Input: convex set S, convex fun. $f : S \to \mathbb{R}$, $x_1 \in S$, step size $\eta > 0$

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

- Update $y_{t+1} = x_t - \eta \nabla f(x_t)$
- Project $x_{t+1} = \arg\min_{x \in S} \|x - y_{t+1}\|$

Ignore project step for the analysis: $x_{t+1} = x_t - \eta \nabla f(x_t)$

We let $v_t = \nabla f(x_t)$, for brevity. Let x^* be **minimizer** of f in S.

Gradient Descent

(Projected) Gradient Descent

Input: convex set S, convex fun. $f : S \to \mathbb{R}$, $x_1 \in S$, step size $\eta > 0$

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

- Update $y_{t+1} = x_t - \eta \nabla f(x_t)$
- Project $x_{t+1} = \arg\min_{x \in S} \|x - y_{t+1}\|$

Ignore project step for the analysis: $x_{t+1} = x_t - \eta \nabla f(x_t)$

We let $v_t = \nabla f(x_t)$, for brevity. Let x^* be **minimizer** of f in S.

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

$$\text{Loss}_{GD} = \sum_{t=1}^{T} (f(x_t) - f(x^*)) \leq \sum_{t=1}^{T} v_t (x_t - x^*) \tag{1}$$

Gradient Descent

(Projected) Gradient Descent

Input: convex set S, convex fun. $f : S \to \mathbb{R}$, $x_1 \in S$, step size $\eta > 0$

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

- Update $y_{t+1} = x_t - \eta \nabla f(x_t)$
- Project $x_{t+1} = \arg\min_{x \in S} \|x - y_{t+1}\|$

Ignore project step for the analysis: $x_{t+1} = x_t - \eta \nabla f(x_t)$

We let $v_t = \nabla f(x_t)$, for brevity. Let x^* be **minimizer** of f in S.

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

$$\text{Loss}_{GD} = \sum_{t=1}^{T} (f(x_t) - f(x^*)) \leq \sum_{t=1}^{T} v_t (x_t - x^*) \tag{1}$$

We show that $\text{Loss}_{GD}/T \to 0$, as $T \to \infty$: the **average** of $f(x_1), \ldots, f(x_T)$ (and $f((x_1 + \cdots + x_T)/T)$ **converges** to optimum $f(x^*)$.

Gradient Descent

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

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

Gradient Descent

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

$$\begin{aligned} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_t - \eta\, \boldsymbol{v}_t - \boldsymbol{x}^*\|^2 \\ &= \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta\, \boldsymbol{v}_t(\boldsymbol{x}_t - \boldsymbol{x}^*) + \eta^2 \|\boldsymbol{v}_t\|^2 \quad \Rightarrow \\ \boldsymbol{v}_t(\boldsymbol{x}_t - \boldsymbol{x}^*) &= \tfrac{1}{2\eta}(\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2) + \tfrac{\eta}{2}\|\boldsymbol{v}_t\|^2 \end{aligned} \tag{2}$$

Gradient Descent

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

$$\begin{aligned} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_t - \eta\, \boldsymbol{v}_t - \boldsymbol{x}^*\|^2 \\ &= \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta\, \boldsymbol{v}_t(\boldsymbol{x}_t - \boldsymbol{x}^*) + \eta^2 \|\boldsymbol{v}_t\|^2 \quad \Rightarrow \\ \boldsymbol{v}_t(\boldsymbol{x}_t - \boldsymbol{x}^*) &= \tfrac{1}{2\eta}(\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2) + \tfrac{\eta}{2}\|\boldsymbol{v}_t\|^2 \end{aligned} \tag{2}$$

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

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

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 $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta \nabla f(\boldsymbol{x}_t)$, we get that:

$$\begin{aligned} \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 &= \|\boldsymbol{x}_t - \eta\, \boldsymbol{v}_t - \boldsymbol{x}^*\|^2 \\ &= \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - 2\eta\, \boldsymbol{v}_t(\boldsymbol{x}_t - \boldsymbol{x}^*) + \eta^2 \|\boldsymbol{v}_t\|^2 \quad \Rightarrow \\ \boldsymbol{v}_t(\boldsymbol{x}_t - \boldsymbol{x}^*) &= \tfrac{1}{2\eta}(\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2) + \tfrac{\eta}{2}\|\boldsymbol{v}_t\|^2 \end{aligned} \tag{2}$$

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

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

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

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

α-Strongly Convex Functions

Using α-strong convexity

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

we get that:

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

α-Strongly Convex Functions

Using α-strong convexity

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

we get that:

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

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

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