Week 1 – L02
Convex Optimization and Gradient Descent (cont)

CS 295 Optimization for Machine Learning

Ioannis Panageas

Analysis of GD for L-smooth, μ-convex

Theorem (Gradient Descent)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, μ-strongly convex (want to minimize) and L-smooth. Let $R = \|x_0 - x^*\|_2$. It holds for $T = \frac{2L}{\mu} \ln\left(\frac{R}{\epsilon}\right)$*

$$\|x_T - x^*\|_2 \leq \epsilon,$$

with appropriately choosing $\alpha = \frac{1}{L}$.

Analysis of GD for L-smooth, μ-convex

Theorem (Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, μ-strongly convex (want to minimize) and L-smooth. Let $R = \|x_0 - x^*\|_2$. It holds for $T = \frac{2L}{\mu} \ln\left(\frac{R}{\epsilon}\right)$*

$$\|x_T - x^*\|_2 \leq \epsilon,$$

with appropriately choosing $\alpha = \frac{1}{L}$.

Proof of Theorem. It holds that

$$\|x_T - x^*\|_2^2 = \left\|x_{T-1} - \frac{1}{L}\nabla f(x_{T-1}) - x^*\right\|_2^2 =$$

Analysis of GD for L-smooth, μ-convex

Theorem (Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, μ-strongly convex (want to minimize) and L-smooth. Let $R = \|x_0 - x^*\|_2$. It holds for $T = \frac{2L}{\mu} \ln\left(\frac{R}{\epsilon}\right)$*

$$\|x_T - x^*\|_2 \leq \epsilon,$$

with appropriately choosing $\alpha = \frac{1}{L}$.

Proof of Theorem. It holds that

$$\|x_T - x^*\|_2^2 = \left\|x_{T-1} - \frac{1}{L}\nabla f(x_{T-1}) - x^*\right\|_2^2 =$$

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

Analysis of GD for L-smooth, μ-convex

Theorem (Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, μ-strongly convex (want to minimize) and L-smooth. Let $R = \|x_0 - x^*\|_2$. It holds for $T = \frac{2L}{\mu} \ln\left(\frac{R}{\epsilon}\right)$*

$$\|x_T - x^*\|_2 \leq \epsilon,$$

with appropriately choosing $\alpha = \frac{1}{L}$.

Proof of Theorem. It holds that

$$\|x_T - x^*\|_2^2 = \left\|x_{T-1} - \frac{1}{L}\nabla f(x_{T-1}) - x^*\right\|_2^2 =$$

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

From Exercise 2 and then Claim 2 we get

$$\frac{2}{L}\nabla f(x_{T-1})^\top (x^* - x_{T-1}) \leq \frac{2}{L}(f(x^*) - f(x_{T-1})) - \frac{\mu}{L}\|x^* - x_{T-1}\|_2^2.$$

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

Analysis of GD for L-smooth, μ-convex

Theorem (Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, μ-strongly convex (want to minimize) and L-smooth. Let $R = \|x_0 - x^*\|_2$. It holds for $T = \frac{2L}{\mu} \ln\left(\frac{R}{\epsilon}\right)$*

$$\|x_T - x^*\|_2 \leq \epsilon,$$

with appropriately choosing $\alpha = \frac{1}{L}$.

Proof of Theorem. It holds that

$$\|x_T - x^*\|_2^2 = \left\|x_{T-1} - \tfrac{1}{L}\nabla f(x_{T-1}) - x^*\right\|_2^2 =$$

$$= \|x_{T-1} - x^*\|_2^2 + \tfrac{1}{L^2}\|\nabla f(x_{T-1})\|_2^2 - 2\tfrac{1}{L}\nabla f(x_{T-1})^\top (x_{T-1} - x^*)$$

Therefore $\|x_T - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)\|x_{T-1} - x^*\|_2^2$.

Analysis of GD for L-smooth, μ-convex

Theorem (Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, μ-strongly convex (want to minimize) and L-smooth. Let $R = \|x_0 - x^*\|_2$. It holds for $T = \frac{2L}{\mu} \ln\left(\frac{R}{\epsilon}\right)$*

$$\|x_T - x^*\|_2 \leq \epsilon,$$

with appropriately choosing $\alpha = \frac{1}{L}$.

Proof of Theorem. It holds that

$$\|x_T - x^*\|_2^2 = \left\|x_{T-1} - \frac{1}{L}\nabla f(x_{T-1}) - x^*\right\|_2^2 =$$

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

Therefore $\|x_T - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)\|x_{T-1} - x^*\|_2^2$.

Thus $\|x_T - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^T R^2 \leq e^{-\frac{\mu T}{L}} R^2$.

Analysis of GD for L-smooth, μ-convex

Theorem (Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, μ-strongly convex (want to minimize) and L-smooth. Let $R = \|x_0 - x^*\|_2$. It holds for $T = \frac{2L}{\mu} \ln\left(\frac{R}{\epsilon}\right)$*

$$\|x_T - x^*\|_2 \leq \epsilon,$$

with appropriately choosing $\alpha = \frac{1}{L}$.

Pr

Remark (last iterate convergence!): $x_T \to x^*$

$$\|x_T - x^*\|_2^2 = \left\|x_{T-1} - \tfrac{1}{L}\nabla f(x_{T-1}) - x^*\right\|_2^2 =$$

$$= \|x_{T-1} - x^*\|_2^2 + \tfrac{1}{L^2}\|\nabla f(x_{T-1})\|_2^2 - 2\tfrac{1}{L}\nabla f(x_{T-1})^\top (x_{T-1} - x^*)$$

Therefore $\|x_T - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)\|x_{T-1} - x^*\|_2^2$.

Thus $\|x_T - x^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^T R^2 \leq e^{-\frac{\mu T}{L}} R^2$.

Projected Gradient Descent (PGD)

(for differentiable functions)

Definition (Projected Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable (want to minimize) in some compact convex set $\mathcal{X}$. The algorithm below is called projected gradient descent*

$$x_{k+1} = \Pi_{\mathcal{X}}(x_k - \alpha \nabla f(x_k)).$$

Remarks

- The projection might not be efficient (is also an optimization problem)!!
- The minimizer x^* does not necessarily satisfy $\nabla f(x^*) = 0$.

Question: When the last remark can be true?

Analysis of Projected GD for L-Lipschitz

Theorem (Projected Gradient Descent). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable, convex (want to minimize in some compact set $\mathcal{X}$) and L-Lipschitz. Let $R = \|x_1 - x^*\|_2$, the distance between the initial point x_0 and minimizer x^*. It holds for $T = \frac{R^2 L^2}{\epsilon^2}$*

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \epsilon,$$

with appropriately choosing $\alpha = \frac{\epsilon}{L^2}$.

Remark

- Same guarantees as in the unconstrained case.

Analysis of Projected GD for L-Lipschitz

Proof. Set $y_t := x_t - \alpha \nabla f(x_t)$. It holds that

$f(x_t) - f(x^*) \leq \nabla f(x_t)^\top (x_t - x^*)$ FOC for convexity,

Analysis of Projected GD for L-Lipschitz

Proof. Set $y_t := x_t - \alpha \nabla f(x_t)$. It holds that

$$
\begin{aligned}
f(x_t) - f(x^*) &\leq \nabla f(x_t)^\top (x_t - x^*) \text{ FOC for convexity,} \\
&= \frac{1}{\alpha}(x_t - y_t)^\top (x_t - x^*) \text{ definition of GD,}
\end{aligned}
$$

Analysis of Projected GD for L-Lipschitz

Proof. Set $y_t := x_t - \alpha \nabla f(x_t)$. It holds that

$$
\begin{aligned}
f(x_t) - f(x^*) &\leq \nabla f(x_t)^\top (x_t - x^*) \text{ FOC for convexity,} \\
&= \frac{1}{\alpha}(x_t - y_t)^\top (x_t - x^*) \text{ definition of GD,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - y_t\|_2^2 - \|y_t - x^*\|_2^2\right) \text{ law of Cosines,}
\end{aligned}
$$

Analysis of Projected GD for L-Lipschitz

Proof. Set $y_t := x_t - \alpha \nabla f(x_t)$. It holds that

$$
\begin{aligned}
f(x_t) - f(x^*) &\le \nabla f(x_t)^\top (x_t - x^*) \text{ FOC for convexity,} \\
&= \frac{1}{\alpha}(x_t - y_t)^\top (x_t - x^*) \text{ definition of GD,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - y_t\|_2^2 - \|y_t - x^*\|_2^2\right) \text{ law of Cosines,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|y_t - x^*\|_2^2\right) + \frac{\alpha}{2}\|\nabla f(x_t)\|_2^2 \text{ Def. of } y_t,
\end{aligned}
$$

Analysis of Projected GD for L-Lipschitz

Proof. Set $y_t := x_t - \alpha \nabla f(x_t)$. It holds that

$$
\begin{aligned}
f(x_t) - f(x^*) &\leq \nabla f(x_t)^\top (x_t - x^*) \text{ FOC for convexity,} \\
&= \frac{1}{\alpha}(x_t - y_t)^\top (x_t - x^*) \text{ definition of GD,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 + \|x_t - y_t\|_2^2 - \|y_t - x^*\|_2^2\right) \text{ law of Cosines,} \\
&= \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|y_t - x^*\|_2^2\right) + \frac{\alpha}{2}\|\nabla f(x_t)\|_2^2 \text{ Def. of } y_t, \\
&\leq \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|y_t - x^*\|_2^2\right) + \frac{\alpha L^2}{2}.
\end{aligned}
$$

Recall. *Suppose $f(x)$ is L-Lipschitz continous.*
Then $\forall x \in dom(f)$

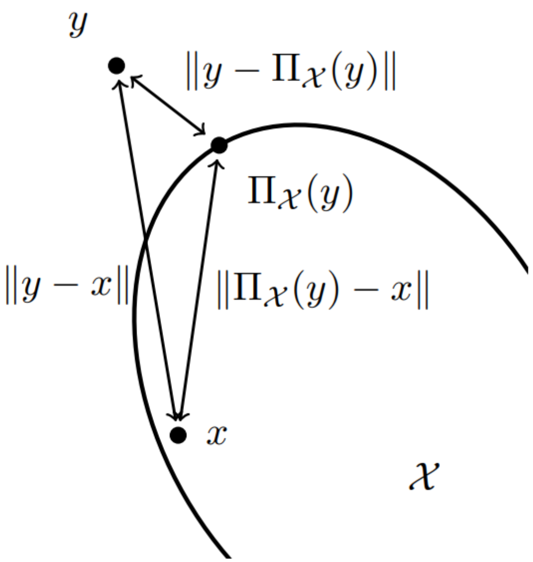
$$\|\nabla f(x)\|_2 \leq L.$$

Analysis of Projected GD for L-Lipschitz

Claim. *It is true that*

$$(\Pi_{\mathcal{X}}(y) - x)^\top (\Pi_{\mathcal{X}}(y) - y) \leq 0.$$

Proof. By picture.



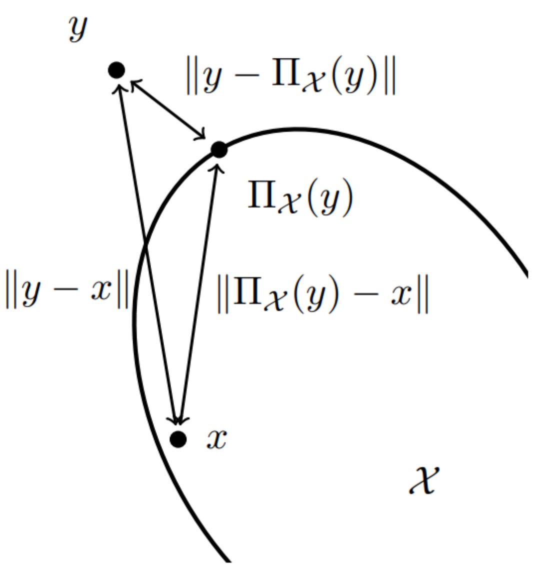
Corollary. *It is true that (Law of Cosines)*

$$\|y - x\|_2^2 \geq \|\Pi_{\mathcal{X}}(y) - y\|_2^2 + \|\Pi_{\mathcal{X}}(y) - x\|_2^2$$

Analysis of Projected GD for L-Lipschitz

$$\text{Therefore } \|y_t - x^*\|_2^2 \geq \|x_{t+1} - y\|_2^2 + \|x_{t+1} - x^*\|_2^2$$
$$\geq \|x_{t+1} - x^*\|_2^2$$

Proof. By picture.



Corollary. *It is true that (Law of Cosines)*

$$\|y - x\|_2^2 \geq \|\Pi_{\mathcal{X}}(y) - y\|_2^2 + \|\Pi_{\mathcal{X}}(y) - x\|_2^2$$

Analysis of Projected GD for L-Lipschitz

Same as in classic GD!

Proof cont. Since

$$f(x_t) - f(x^*) \leq \frac{1}{2\alpha}\left(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2\right) + \frac{\alpha L^2}{2},$$

taking the telescopic sum we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) &\leq \frac{1}{2\alpha T}(\|x_1 - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2) + \frac{\alpha L^2}{2}.\\ &\leq \frac{R^2}{2\alpha T} + \frac{\alpha L^2}{2} = \epsilon \text{ by choosing appropriately } \alpha, T.\end{aligned}$$

The claim follows by convexity since $\frac{1}{T}\sum_{t=1}^{T} f(x_t) \geq f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)$ (Jensen's inequality).

Conclusion

- Introduction to Convex Optimization
  - Easy to minimize (generally is NP-hard).
  - GD has rate of convergence $\mathrm{O}\left(\frac{L^2}{\epsilon^2}\right)$ for L-Lipschitz.
  - GD has rate of convergence $\mathrm{O}\left(\frac{L}{\epsilon}\right)$ for L-smooth.
  - GD has rate of convergence $\mathrm{O}\left(\frac{L}{\mu}\ln\frac{1}{\epsilon}\right)$ for L-smooth, μ-convex.
  - Same is true for *Projected* GD (similar analysis) for constrained optimization.
- Next week we will talk about sub-gradients (non-differentiable functions) and Stochastic Gradient Descent (SGD).