SDSC6015 Stochastic Optimization for Machine Learning

Lu Yu

Department of Data Science, City University of Hong Kong

September 18, 2025

Gradient Descent and Subgradient Method

Convex Optimization Problems

$$\min_{x \in \mathbb{R}^d} f(x),$$

where

- f is a **convex** function
- $\mathbb{R}^d$ is convex
- $\boldsymbol{x}^*$ is the minimizer of function f :

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})$$

Recap: Gradient Descent

Update rule for gradient descent:

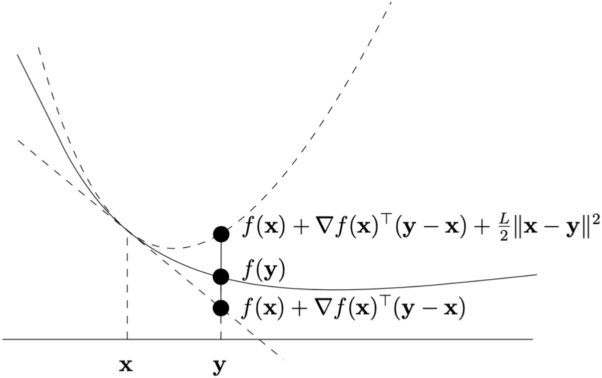
$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_{k+1} \nabla f(\boldsymbol{x}_k)$$

- $\boldsymbol{x}_k$: current point (parameters or variables).
- η_k : step size (learning rate), a positive scalar determining how far we move in the gradient direction.
- $\boldsymbol{x}_{k+1}$: next point after the update.

Definition

Let $f : \textbf{dom}(f) \to \mathbb{R}$ be differentiable, $X \subseteq \textbf{dom}(f)$, $L > 0$. f is called smooth (with parameter L) over X if

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X.$$



Subgradient

Recall: for convex and differentiable $f : \mathbb{R}^d \to \mathbb{R}$

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \qquad \forall \boldsymbol{x}, \boldsymbol{y}.$$

Definition

A subgradient of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at $\boldsymbol{x}$ is any $g \in \mathbb{R}^d$ such that

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + g^\top (\boldsymbol{y} - \boldsymbol{x}), \qquad \forall \boldsymbol{y}\,.$$

Set of all subgradients of f is called the subdifferential:

$$\partial f(\boldsymbol{x}) = \{g \in \mathbb{R}^d : g \text{ is a subgradient of } f \text{ at } \boldsymbol{x}\}\,.$$

Subgradient Method

Now consider convex function $f : \mathbb{R}^d \to \mathbb{R}$ convex, but not necessarily differentiable.

Subgradient method: like gradient descent, but replacing gradients with subgradients

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_{k+1} g_k$$

- $\boldsymbol{x}_k$: current point
- $g_k \in \nabla f(\boldsymbol{x}_k)$: any subgradient of f at $\boldsymbol{x}_k$
- $\eta_k > 0$: step size
- $\boldsymbol{x}_{k+1}$: next point after the update.

Caveat: Subgradient method is not necessarily a descent method!
e.g. $f(x) = |x|$ (non-smoothness causes oscillation)

Summary

f	Algorithm	Convergence	# Iterations
Convex L-Lipschitz	GD	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{RL}{\sqrt{T}}$	$\frac{R^2L^2}{\varepsilon^2}$
Convex L-Smooth	GD	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{R^2L}{2T}$	$\frac{R^2L}{2\varepsilon}$
Convex L-Lipschitz	Subgrad	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{LR}{\sqrt{T}}$	$\frac{R^2L^2}{\varepsilon^2}$

- Time horizon $T > 0$ is given
- $R := \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$
- $\boldsymbol{x}_{\text{best}}^{(T)} := \arg\min_{i=0,1,\ldots,T} f(\boldsymbol{x}_i)$.

Thus, the subgradient method has convergence rate $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$
...compare this to $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ rate of gradient descent

Faster Gradient Descent

Can we go even faster?

- So far: Error decreases with $1/\sqrt{T}$, or $1/T$...
- Could the error decrease exponentially in T, i.e.,

$$e^{-cT}, \quad \text{for some } c > 0,$$

rather than following a polynomial decay[1]?

[1] Polynomial decay refers to a rate of decrease that follows a power law, meaning a quantity $f(t)$ diminishes over time t as $f(t) = \mathcal{O}(t^{-\alpha})$ for some exponent $\alpha > 0$

Can we go even faster?

- Consider $f(x) = x^2$: step size $\eta = \frac{1}{2}$ (f is $L = 2$ - smooth)

$$x_{t+1} = x_t - \frac{1}{2}\nabla f(x_t) = x_t - x_t = 0$$

Converge in one step!

- Same $f(x) = x^2$: step size $\eta = \frac{1}{4}$

$$x_{t+1} = x_t - \frac{1}{4}\nabla f(x_t) = x_t - \frac{x_t}{2} = \frac{x_t}{2}$$

so

$$f(x_t) = f\Big(\frac{x_0}{2^t}\Big) = \frac{1}{2^{2t}} x_0^2$$

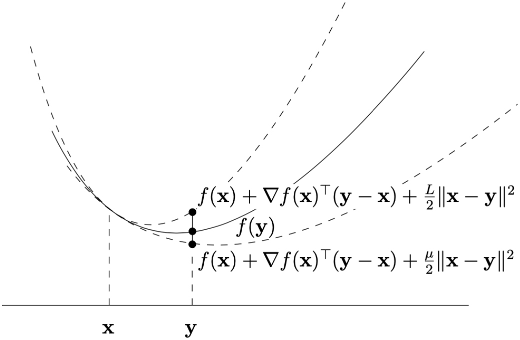
Exponential in t!

Strongly Convex Functions

Definition

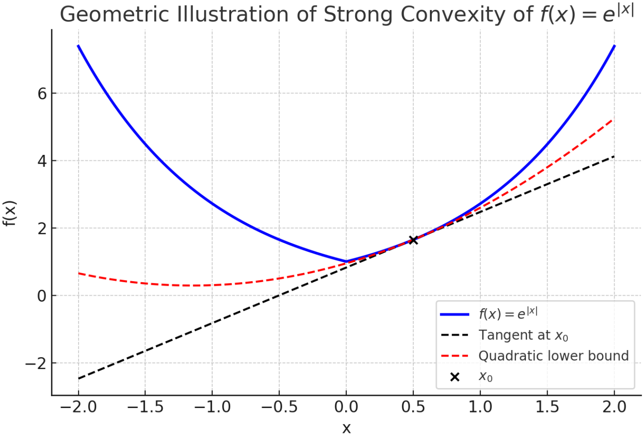
Let f : $\mathbf{dom}(f) \to \mathbb{R}$ be a differentiable function, $\mathbf{dom}(f)$ is a convex set and a constant $\mu > 0$. f is called strongly convex with parameter μ if

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \ \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbf{dom}(f)\,.$$



For any $\boldsymbol{x}$, the graph of f lies above a tangent paraboloid at $(\boldsymbol{x}, f(\boldsymbol{x}))$.

Example: $f(x) = e^{|x|}$ is strongly convex with parameter $\mu = 1$.



Strongly Convex Functions

Strong convexity enforces that the function grows at least a quadratic rate as you move away from the minimum.
This ensures:

- A unique minimizer – there is only one global minimum.

  Lemma 1
  If f is strongly convex, then f is strictly convex and has a unique global minimum.

- Faster optimization convergence – see later!

GD on Smooth and Strongly Convex Functions

- Aim to show $\lim_{t\to\infty} \boldsymbol{x}_t = \boldsymbol{x}^*$.

Step 1: Vanilla Analysis from the last lecture:

$$\nabla f(\boldsymbol{x}_t)^\top(\boldsymbol{x}_t - \boldsymbol{x}^*) = \frac{\eta}{2}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{1}{2\eta}(\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2)$$

Step 2: Now use stronger lower bound on the left hand side, coming from strong convexity

$$\nabla f(\boldsymbol{x}_t)^\top(\boldsymbol{x}_t - \boldsymbol{x}^*) \geqslant f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) + \frac{\mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2$$

Step 3: Putting it together and rearranging gives

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant 2\eta\big(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)\big) + \eta^2\|\nabla f(\boldsymbol{x}_t)\|^2 + (1-\mu\eta)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2$$

GD on Smooth and Strongly Convex Functions

$$\underline{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2} \leqslant \underbrace{2\eta\big(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)\big) + \eta^2\|\nabla f(\boldsymbol{x}_t)\|^2}_{\text{“noise”}} + \underline{(1 - \mu\eta)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2}$$

- Choose $\eta < \frac{1}{\mu}$
- Squared distance to $\boldsymbol{x}^*$ goes down by a constant factor $(1 - \mu\eta)$, up to some “noise”

GD on Smooth and Strongly Convex Functions

Theorem 1

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $\boldsymbol{x}^*$; suppose that f is L-smooth and strongly convex with parameter μ. Choosing

$$\eta = \frac{1}{L},$$

gradient descent with arbitrary $\boldsymbol{x}_0$ satisfies

- Squared distances to $\boldsymbol{x}^*$ are geometrically decreasing:

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant \left(1 - \frac{\mu}{L}\right)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2, \qquad t > 0\,.$$

- The absolute error after T iterations is exponentially with T:

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \qquad T > 0.$$

GD on Smooth and Strongly Convex Functions

$$\underline{\|\boldsymbol{x}_{t+1}-\boldsymbol{x}^*\|^2} \leqslant \underbrace{2\eta\big(f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t)\big)+\eta^2\|\nabla f(\boldsymbol{x}_t)\|^2}_{\text{“noise”}}+\underline{(1-\mu\eta)\|\boldsymbol{x}_t-\boldsymbol{x}^*\|^2}$$

Proof of (i).

Bounding the noise Note that $\eta = 1/L$

$$\begin{aligned}2\eta\big(f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t)\big)+\eta^2\|\nabla f(\boldsymbol{x}_t)\|^2 &= \frac{2}{L}\big(f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t)\big)+\frac{1}{L^2}\|\nabla f(\boldsymbol{x}_t)\|^2 \\ &\leqslant \frac{2}{L}\big(f(\boldsymbol{x}_{t+1})-f(\boldsymbol{x}_t)\big)+\frac{1}{L^2}\|\nabla f(\boldsymbol{x}_t)\|^2\end{aligned}$$

Employing Lemma 3 (sufficient decrease) from the last lecture:

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

Hence, the noise is nonpositive

$$2\eta\big(f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t)\big)+\eta^2\|\nabla f(\boldsymbol{x}_t)\|^2 \leqslant -\frac{1}{L^2}\|\nabla f(\boldsymbol{x}_t)\|^2+\frac{1}{L^2}\|\nabla f(\boldsymbol{x}_t)\|^2 = 0\,.$$

GD on Smooth and Strongly Convex Functions

Then, we get (i):

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 \leqslant (1 - \mu\eta)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 = \left(1 - \frac{\mu}{L}\right)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2$$

Proof of (ii). From (i):

$$\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \leqslant \left(1 - \frac{\mu}{L}\right)^T \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 .$$

Smoothness together with $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$:

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{x}_T - \boldsymbol{x}^*) + \frac{L}{2}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 = \frac{L}{2}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 .$$

Putting it together:

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \frac{L}{2}\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 \leqslant \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 .$$

GD on Smooth and Strongly Convex Functions

$R^2 := \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2$

$$T \geqslant \frac{L}{\mu} \ln\left(\frac{R^2 L}{2\varepsilon}\right) \quad \Rightarrow \quad \text{error } \leqslant \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T R^2 \leqslant \varepsilon$$

Conclusion: To reach absolute error at most ε, we only need $\mathcal{O}\big(\log \frac{1}{\varepsilon}\big)$ iterations, e.g.

- $\frac{L}{\mu} \ln(50 \cdot R^2 L)$ iterations for error $\varepsilon = 0.01$
- ... as opposed to $50 \cdot R^2 L$ in the smooth case

In Practice:

What if we don't know the smoothness parameter L?

Questions?

Summary

f	Algorithm	Convergence	# Iterations
Convex L-Lipschitz	GD	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{RL}{\sqrt{T}}$	$\frac{R^2L^2}{\varepsilon^2}$
Convex L-Smooth	GD	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{R^2L}{2T}$	$\frac{R^2L}{2\varepsilon}$
μ-Strongly Convex L-Smooth	GD	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{RL}{2}(1 - \frac{\mu}{L})^T$	$\frac{L}{\mu} \ln \left(\frac{R^2L}{2\varepsilon} \right)$
Convex L-Lipschitz	Subgrad	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{LR}{\sqrt{T}}$	$\frac{R^2L^2}{\varepsilon^2}$

Faster Subgradient Method

Last time

Given a convex and L-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ (f is not necessarily differentiable), the subgraident method with $R = \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$ satisfies

$$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{LR}{\sqrt{T}}$$

Question: Can we improve the convergence rate?

Optimality of First-order Methods

With all the convergence rates we have seen so far, a very natural question to ask is if these rates are the best possible or not. Surprisingly, the rate can indeed not be improved in general.

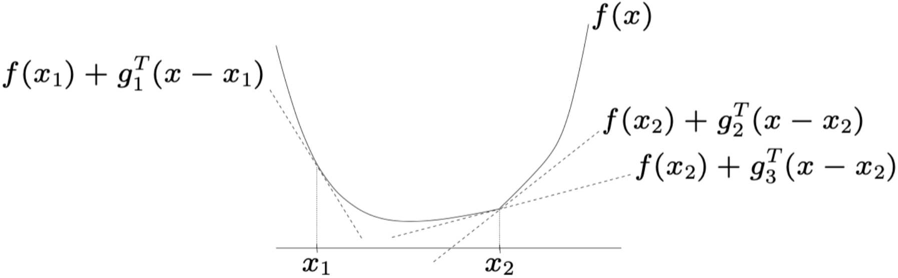
Theorem 2

For any $T \leqslant d - 1$ and starting point $\boldsymbol{x}_0$, there is a function f in the problem class of L-Lipschitz functions over $\mathbb{R}^d$, such that any (sub)gradient method has an objective error at least

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \geqslant \frac{LR}{2(1 + \sqrt{T + 1})}$$

Smooth (non-differentiable) functions?

- They don't exist: A non-differentiable function cannot be smooth (e.g. $f(x) = |x|$).
- Can we still improve over $\mathcal{O}(1/\varepsilon^2)$ steps for Lipschitz functions?
- Yes, if we also require strong convexity.



Straightforward generalization to the non-differentiable case:

Definition.
Let $f : \textbf{dom}(f) \to \mathbb{R}$ be convex, and $\mu > 0$. Function f is called strongly convex with parameter μ if

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \boldsymbol{g}^\top(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \textbf{dom}(f), \forall \boldsymbol{g} \in \partial f(\boldsymbol{x})$$

Does strong convexity alone guarantee a fast convergence?

NO!

Reason: The subgradient method does not fully exploit strong convexity, since subgradients

- are not always well-aligned with the optimal descent direction
- lack curvature information (e.g. Hessian information)

Strongly Convex Functions

For fast convergence, we consider additional assumptions beyond strongly convexity.

- Smoothness? - Not an option in the non-differentiable case.
- Instead: assume that all subgradients $\boldsymbol{g}_t$ during the algorithm are bounded in norm.
  - This is not always equivalent to Lipschitz (see notes)
  - Over $\mathbb{R}^d$, strong convexity and Lipschitz continuity contradict each other (see notes)

Subgradient Method on Strongly Convex Functions

Theorem 3

Let $f : \mathbb{R}^d \to \mathbb{R}$ be strongly convex with parameter $\mu > 0$ and let $\boldsymbol{x}^*$ be the unique global minimum of f. With decreasing step size

$$\eta_t = \frac{2}{\mu(t+1)}, \quad t > 0,$$

subgradient descent yields

$$\underbrace{f\left(\frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \boldsymbol{x}_t\right)}_{\text{convex combination of iterates}} - f(\boldsymbol{x}^*) \leqslant \frac{2B^2}{\mu(T+1)},$$

where $B = \max_{t=1}^{T} \|\boldsymbol{g}_t\|$.

Proof.

Vanilla analysis: for any $\boldsymbol{g}_t \in \partial f(\boldsymbol{x}_t)$

$$\boldsymbol{g}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}^*) = \frac{\eta_t}{2}\|\boldsymbol{g}_t\|^2 + \frac{1}{2\eta_t}\left(\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2\right).$$

Lower bound from strong convexity:

$$\boldsymbol{g}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}^*) \geqslant f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) + \frac{\mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2.$$

Putting it together with $\|\boldsymbol{g}_t\|^2 \leqslant B^2$:

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leqslant \frac{B^2\eta_t}{2} + \frac{\eta_t^{-1} - \mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \frac{\eta_t^{-1}}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2.$$

Summing over $t = 1, \ldots, T$: we used to have telescoping with $\eta_t = \eta, \mu = 0$ in the previous

Subgradient Method on Strongly Convex Functions

So far we have:

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leqslant \frac{B^2\eta_t}{2} + \frac{\eta_t^{-1} - \mu}{2}\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \frac{\eta_t^{-1}}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2 .$$

Plug in $\eta_t^{-1} = \mu(1+t)/2$ multiply with t on both sides:

$$\begin{aligned} t \cdot \big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)\big) &\leqslant \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\Big(t(t-1)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - t(t+1)\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|\Big) \\ &\leqslant \frac{B^2}{\mu} + \frac{\mu}{4}\Big(t(t-1)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - t(t+1)\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2\Big) \end{aligned}$$

Now we get telescoping

$$\sum_{t=1}^{T} t \cdot \big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)\big) \leqslant \frac{TB^2}{\mu} + \frac{\mu}{4}\Big(0 - T(T+1)\|\boldsymbol{x}_{T+1} - \boldsymbol{x}^*\|^2\Big) \leqslant \frac{TB^2}{\mu}\,.$$

Since

$$\frac{2}{T(T+1)} \sum_{t=1}^{T} t = 1\,,$$

Jensen's inequality yields

$$f\Big(\frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \boldsymbol{x}_t\Big) - f(\boldsymbol{x}^*) \leqslant \frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \big(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)\big)\,.$$

Subgradient Method on Strongly Convex Functions

Putting together

$$f\Big(\frac{2}{T(T+1)}\sum_{t=1}^{T} t \cdot \boldsymbol{x}_t\Big) - f(\boldsymbol{x}^*) \leqslant \frac{2B^2}{\mu(T+1)} \,.$$

- Weighted average of iterates achieves the bound (later iterates have more weight)
- Bound is independent of initial distance $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$?
    - Not really: B typically depends on $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$
    - for example, $B = \mathcal{O}(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|)$ for quadratic functions

Subgradient Method on Strongly Convex Functions

$$T \geqslant \frac{2B^2}{\mu\varepsilon} \quad \Rightarrow \quad \text{error } \leqslant \frac{2B^2}{\mu T} \leqslant \varepsilon$$

Conclusion: To reach absolute error at most ε, we need $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ iterations,

- Recall: we can only hope that B is small (can be checked while running the algorithm)
- What if we don't know the parameter μ of strong convexity?
  - Heuristic strategy: try some μ's, pick best solution obtained
  - Choosing the step size without μ...

Questions?

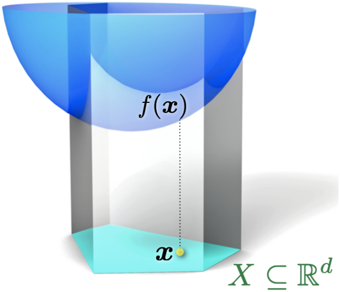
f	Algorithm	Convergence	# Iterations
Convex L-Lipschitz	GD	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{RL}{\sqrt{T}}$	$\frac{R^2L^2}{\varepsilon^2}$
Convex L-Smooth	GD	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{R^2L}{2T}$	$\frac{R^2L}{2\varepsilon}$
μ-Strongly Convex L-Smooth	GD	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{RL}{2}(1 - \frac{\mu}{L})^T$	$\frac{L}{\mu} \ln \left(\frac{R^2L}{2\varepsilon}\right)$
Convex L-Lipschitz	Subgrad	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{LR}{\sqrt{T}}$	$\frac{R^2L^2}{\varepsilon^2}$
μ-Strongly Convex $\Vert\boldsymbol{g}\Vert \leqslant B$	Subgrad	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{2B^2}{\mu(T+1)}$	$\frac{2B^2}{\mu\varepsilon}$

Here, $R = \Vert\boldsymbol{x}_0 - \boldsymbol{x}^*\Vert$

Projected Gradient Descent

Constrained Optimization Problem

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in X \end{aligned}$$



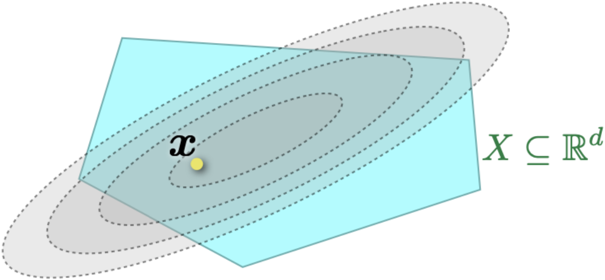
Constrained Optimization

Solving Constrained Optimization Problem

$$\begin{aligned} \text{minimize} \quad & f(\boldsymbol{x}) \\ \text{subject to} \quad & \boldsymbol{x} \in X \end{aligned}$$

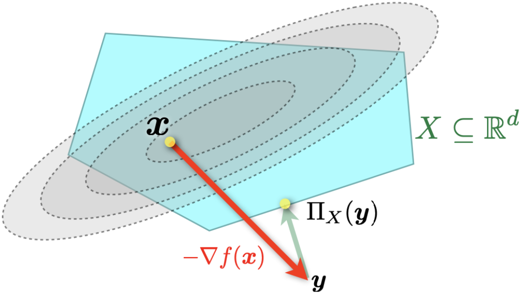
Solution:

- Projected Gradient Descent
- Transform it into an *unconstrained* problem



Projected Gradient Descent

Idea: project onto X after every step: $\Pi_X(\boldsymbol{y}) := \arg\min_{\boldsymbol{x}\in X} \|\boldsymbol{x} - \boldsymbol{y}\|$



Projected gradient descent: $\boldsymbol{x}_{t+1} = \Pi_X[\boldsymbol{x}_t - \eta\nabla f(\boldsymbol{x}_t)]$

Projected Gradient Descent

Projected gradient descent:

$$\boldsymbol{y}_{t+1} := \boldsymbol{x}_t - \eta_t \nabla f(\boldsymbol{x}_t)$$
$$\boldsymbol{x}_{t+1} := \Pi_X(\boldsymbol{y}_{t+1}) = \operatorname*{arg\,min}_{\boldsymbol{x} \in X} \|\boldsymbol{x} - \boldsymbol{y}\|$$

for stepsize $\eta_t > 0$ and timesteps $t = 0, 1, \ldots$

Properties of Projection

Fact 1.

Let $X \subseteq \mathbb{R}^d$ be closed and convex, $\boldsymbol{x} \in X, \boldsymbol{y} \in \mathbb{R}^d$. Then

- $(\boldsymbol{x} - \Pi_X(\boldsymbol{y}))^\top (\boldsymbol{y} - \Pi_X(\boldsymbol{y})) \leqslant 0$
- $\|\boldsymbol{x} - \Pi_X(\boldsymbol{y})\|^2 + \|\boldsymbol{y} - \Pi_X(\boldsymbol{y})\|^2 \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|^2$

Proof.

(i) $\Pi_X(\boldsymbol{y})$ is the minimizer of (differentiable) convex function $d_{\boldsymbol{y}} = \|\boldsymbol{x} - \boldsymbol{y}\|^2$ over X.
By first-order characterization of optimality (Lemma 4 from Lecture 2),

$$
\begin{array}{cccl}
 & 0 & \leqslant & \nabla d_{\boldsymbol{y}}(\Pi_X(\boldsymbol{y}))^\top (\boldsymbol{x} - \Pi_X(\boldsymbol{y})) \\
 & & = & 2(\Pi_X(\boldsymbol{y}) - \boldsymbol{y})^\top (\boldsymbol{x} - \Pi_X(\boldsymbol{y})) \\
\Leftrightarrow & 0 & \geqslant & 2(\boldsymbol{y} - \Pi_X(\boldsymbol{y}))^\top (\boldsymbol{x} - \Pi_X(\boldsymbol{y})) \\
\Leftrightarrow & 0 & \geqslant & (\boldsymbol{x} - \Pi_X(\boldsymbol{y}))^\top (\boldsymbol{y} - \Pi_X(\boldsymbol{y}))
\end{array}
$$

Let $X \subseteq \mathbb{R}^d$ be closed and convex, $\boldsymbol{x} \in X, \boldsymbol{y} \in \mathbb{R}^d$. Then

- $(\boldsymbol{x} - \Pi_X(\boldsymbol{y}))^\top (\boldsymbol{y} - \Pi_X(\boldsymbol{y})) \leqslant 0$
- $\|\boldsymbol{x} - \Pi_X(\boldsymbol{y})\|^2 + \|\boldsymbol{y} - \Pi_X(\boldsymbol{y})\|^2 \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|^2$

Proof.
(ii)

$$\boldsymbol{v} := (\boldsymbol{x} - \Pi_X(\boldsymbol{y})), \qquad \boldsymbol{w} := (\boldsymbol{y} - \Pi_X(\boldsymbol{y}))\,.$$

By (i),

$$\begin{aligned} 0 \geqslant 2\boldsymbol{v}^\top \boldsymbol{w} &= \|\boldsymbol{v}\|^2 + \|\boldsymbol{w}\|^2 - \|\boldsymbol{v} - \boldsymbol{w}\|^2 \\ &= \|\boldsymbol{x} - \Pi_X(\boldsymbol{y})\|^2 + \|\boldsymbol{y} - \Pi_X(\boldsymbol{y})\|^2 - \|\boldsymbol{x} - \boldsymbol{y}\|^2\,. \end{aligned}$$

Convergence Rate of Projected Gradient Descent

The same number of steps as a gradient over $\mathbb{R}^d$!

- Lipschitz convex functions over X : $\mathcal{O}(1/\varepsilon^2)$ steps
- Smooth convex functions over X : $\mathcal{O}(1/\varepsilon)$ steps
- Smooth and strongly convex functions over X : $\mathcal{O}(\log(1/\varepsilon))$ steps

We will adapt the previous proofs for gradient descent.
BUT:

- Each step involves a projection onto X
- may or may not be efficient. . .

Projected GD on Lipschitz Convex Functions

Assume that all gradients of f are bounded in norm over closed and convex X.

- Equivalent to f being Lipschitz over X
- Many interesting functions are Lipschitz over bounded sets X.

Theorem 4 (same as the unconstrained one, but more useful)
Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $X \subseteq \mathbb{R}^d$ closed and convex, $\boldsymbol{x}^*$ is the minimizer of f over X; furthermore, suppose that $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leqslant R$ with $\boldsymbol{x}_0 \in X$, and that $\|\nabla f(\boldsymbol{x})\| \leqslant B$ for all $\boldsymbol{x} \in X$. Choosing the constant stepsize

$$\eta = \frac{R}{B\sqrt{T}},$$

projected gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leqslant \frac{RB}{\sqrt{T}}.$$

Proof.

- Replace $\boldsymbol{x}_{t+1}$ in the vanilla analysis with $\boldsymbol{y}_{t+1}$ (the unprojected gradient step):

$$\boldsymbol{g}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}^*) = \frac{1}{2\eta}\left(\eta^2\|\boldsymbol{g}_t\|^2 + \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{x}^*\|^2\right).$$

- Use Fact 1 (ii): $\|\boldsymbol{x} - \Pi_X(\boldsymbol{y})\|^2 + \|\boldsymbol{y} - \Pi_X(\boldsymbol{y})\|^2 \leqslant \|\boldsymbol{x} - \boldsymbol{y}\|^2$
- With $\boldsymbol{x} = \boldsymbol{x}^*, \boldsymbol{y} = \boldsymbol{y}_{t+1}$ we have $\Pi_X(\boldsymbol{y}) = \boldsymbol{x}_{t+1}$, and hence

$$\|\boldsymbol{x}^* - \boldsymbol{x}_{t+1}\| \leqslant \|\boldsymbol{x}^* - \boldsymbol{y}_{t+1}\|^2$$

- We go back to the original vanilla analysis and continue from there as before:

$$\boldsymbol{g}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}^*) \leqslant \frac{1}{2\eta}\left(\eta^2\|\boldsymbol{g}_t\|^2 + \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2\right)$$

Smooth functions over X

Recall:
f is called smooth (with parameter L) over X if

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \quad \forall \boldsymbol{x}, \boldsymbol{y} \in X.$$

Sufficient Decrease

Lemma 2

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L over X. Choosing stepsize

$$\eta = \frac{1}{L},$$

projected gradient descent with arbitrary $\boldsymbol{x}_0 \in X$ satisfies

$$f(\boldsymbol{x}_{t+1}) \leqslant f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2, \qquad t \geqslant 0.$$

Sufficient Decrease

$$f(\boldsymbol{x}_{t+1}) \leqslant f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2, \qquad t \geqslant 0\,.$$

Proof.

Use smoothness, $\boldsymbol{y}_{t+1} - \boldsymbol{x}_t = -\nabla f(\boldsymbol{x}_t)/L, 2\boldsymbol{v}^\top \boldsymbol{w} = \|\boldsymbol{v}\|^2 + \|\boldsymbol{w}\|^2 - \|\boldsymbol{v} - \boldsymbol{w}\|^2$:

$$\begin{aligned}
f(\boldsymbol{x}_{t+1}) &\leqslant f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)^\top(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 \\
&= f(\boldsymbol{x}_t) - L(\boldsymbol{y}_{t+1} - \boldsymbol{x}_t)^\top(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 \\
&= f(\boldsymbol{x}_t) - \frac{L}{2}\left(\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_t\|^2 + \|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|^2 - \|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2\right) \\
&\qquad + \frac{L}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 \\
&= f(\boldsymbol{x}_t) - \frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_t\|^2 + \frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|^2 \\
&= f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{L}{2}\|\boldsymbol{y}_{t+1} - \boldsymbol{x}_{t+1}\|\,.
\end{aligned}$$

Questions?

Assignment 1 will be released after this class and is due on **October 1 at 11:59 PM**.

References

- Stephen P Boyd, *Lecture notes for ee 264b,stanford university (2010-2011)*.
- Sébastien Bubeck, *Convex optimization: Algorithms and complexity*, Foundations and Trends in Machine Learning **8** (2015), no. 3-4, 231–357.
- Stephen P Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.
- R. Tyrrell Rockafellar, *Convex analysis*, Princeton University Press.