Mathematics of Image and Data Analysis
Math 5467

Lecture 25: Gradient Descent

Instructor: Jeff Calder
Email: jcalder@umn.edu

http://www-users.math.umn.edu/~jwcalder/5467S21

Announcements

- HW4 due April 30, Project 3 due May 9.
- Please fill out *Student Rating of Teaching (SRT)* online as soon as possible, and before **May 3**.
  - You should have received an email from Office of Measurement Services with a link.
  - You can also find a link on our Canvas website.

Last time

- Universal approximation
- Convolutional neural networks

Today

- Gradient Descent

Gradient Descent

Gradient descent is one of the most important algorithms in many areas of science and engineering. To minimize an objective function $f : \mathbb{R}^n \to \mathbb{R}$, gradient descent iterates

$$x_{k+1} = x_k - \alpha \nabla f(x_k) \tag{1}$$

until convergence. The parameter $\alpha > 0$ is the time step (often called the *learning rate* when using gradient descent to train machine learning algorithms).

Assumptions on f

We assume the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is a smooth function that admits a global minimizer $x_* \in \mathbb{R}^n$. That is

$$f(x_*) \leq f(x)$$

for all $x \in \mathbb{R}^n$. We denote the optimal value of f by $f_* := f(x_*)$.

Sublinear convergence rate

We say ∇f is *L-Lipschitz continuous* if

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n. \tag{2}$$

$$f_* = \min_{x \in \mathbb{R}^n} f(x)$$

Theorem 1. *Assume ∇f is L-Lipschitz and that $\alpha \leq \frac{1}{L}$. Then for any integer $t \geq 1$ we have*

$$\min_{0 \leq k \leq t} \|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f_*)}{\alpha t}. \tag{3}$$

$$= O\left(\frac{1}{t}\right)$$

Remark 2. The theorem says, with very few assumptions on f, that gradient descent converges at a rate of $O\left(\frac{1}{t}\right)$ to a critical point of f, in the sense that $\nabla f \sim \frac{1}{t} \to 0$. Since f is not assumed to be convex, critical points need not be minimizers and could be also include saddle points.

Proof: Claim that

One-sided Taylor expansion

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} \|x-y\|^2$$

To see this

Fundamental Theorem of Calc.

$$f(y) - f(x) = \int_0^1 \frac{d}{dt} f(ty + (1-t)x)\, dt$$

$$= \int_0^1 \nabla f(ty + (1-t)x)^T \frac{d}{dt}(ty + (1-t)x)\, dt$$

$$= \int_0^1 \nabla f(ty + (1-t)x)^T (y-x)\, dt$$

$$= \int_0^1 \nabla f(x)^T (y-x) + \left(\nabla f(ty+(1-t)x) - \nabla f(x)\right)^T (y-x)\,dt$$

$$= \nabla f(x)^T (y-x) + \int_0^1 \left(\nabla f(ty+(1-t)x) - \nabla f(x)\right)^T (y-x)\,dt$$

Cauchy-Schwarz

$$x^T y \le \|x\|\,\|y\|$$

Hence

$$f(y) - f(x) \le \nabla f(x)^T (y-x)$$

$$+ \int_0^1 \underbrace{\|\nabla f(ty+(1-t)x) - \nabla f(x)\|}_{\le L\|ty+(1-t)x-x\|}\,\|y-x\|\,dt$$

by L-Lipschitz property

$$= L \| t(y-x) \|$$

$$= L t \| y-x \|$$

$$f(y) - f(x) \le \nabla f(x)^T (y-x) + L \|y-x\|^2 \int_0^1 t \, dt$$

$$= \nabla f(x)^T (y-x) + \frac{L}{2} \|y-x\|^2.$$

Which proves the claim.

Take $y = x_{k+1}$, $x = x_k$, $x_{k+1} = x_k - \alpha \nabla f(x_k)$

$$f(x_{k+1}) - f(x_k) \le \nabla f(x_k)^T \underbrace{(x_{k+1} - x_k)} + \frac{L}{2} \| \underbrace{x_{k+1} - x_k} \|^2$$

$$= -\alpha \nabla f(x_k)$$

$$= \nabla f(x_k)^T \left(-\alpha \nabla f(x_k)\right) + \frac{L}{2} \|-\alpha \nabla f(x_k)\|^2$$

$$= -\alpha \|\nabla f(x_k)\|^2 + \frac{L\alpha^2}{2} \|\nabla f(x_k)\|^2$$

$$= -\underbrace{\left(\alpha - \frac{L\alpha^2}{2}\right)}_{\text{want} \geq 0} \|\nabla f(x_k)\|^2$$

$$\Longrightarrow \alpha \geq \frac{L\alpha^2}{2}$$

Can maximize

or $\alpha \leq \frac{2}{L}$.

$$\alpha - \frac{L\alpha^2}{2} \text{ over } \alpha.$$

$$0 = \frac{d}{d\alpha}\left(\alpha - \frac{L\alpha^2}{2}\right) = 1 - L\alpha$$

$1 - L\alpha = 0$ when $\alpha = \frac{1}{L}$.

Assume $\alpha \le \frac{1}{L}$. In this case

$$\alpha - \frac{L\alpha^2}{2} = \alpha - \frac{L\alpha}{2} \cdot \alpha \quad \alpha \le \frac{1}{L}$$

$$\ge \alpha - \frac{L\alpha}{2} \cdot \frac{1}{L}$$

$$= \alpha - \frac{\alpha}{2} = \frac{\alpha}{2}.$$

Hence if $\alpha \leq \frac{1}{L}$ then

$$\boxed{f(x_{k+1}) \leq f(x_k) - \frac{\alpha}{2}\|\nabla f(x_k)\|^2}$$

Rearrange to get

$$\frac{\alpha}{2}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})$$

telescoping sum.

$$\frac{\alpha}{2}\sum_{k=0}^{t}\|\nabla f(x_k)\|^2 \leq \sum_{k=0}^{t}\left(f(x_k) - f(x_{k+1})\right)$$

$$= f(x_0) - f(x_{t+1})$$

$$\leq f(x_0) - f_*$$

Use bound $\sum_{k=0}^{t} \|\nabla f(x_k)\|^2 \geq \min_{0 \leq k \leq t} \|\nabla f(x_k)\|^2 (t+1)$

$$\min_{0 \leq k \leq t} \|\nabla f(x_k)\|^2 \leq \frac{1}{t+1} \sum_{k=0}^{t} \|\nabla f(x_k)\|^2$$

$$\leq \frac{2}{\alpha(t+1)} \left(f(x_0) - f_*\right) \quad \text{V/I}$$

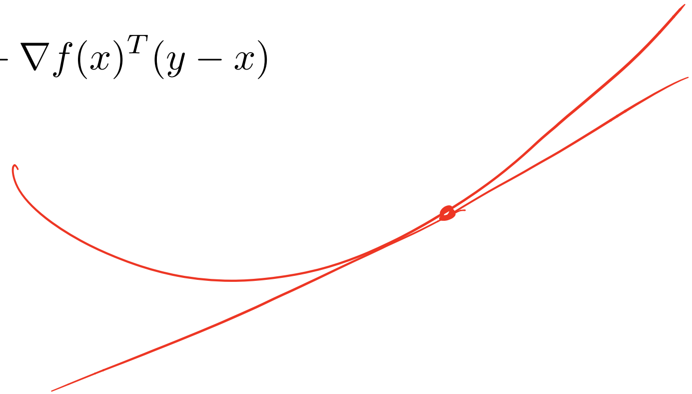
$$\frac{1}{t+1} \leq \frac{1}{t}$$

Convergence to a minimizer

To show that gradient descent converges to a global minimizer of f, we need to assume that f is *convex*, which for us means that f lies above its tangent planes, that is

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{4}$$

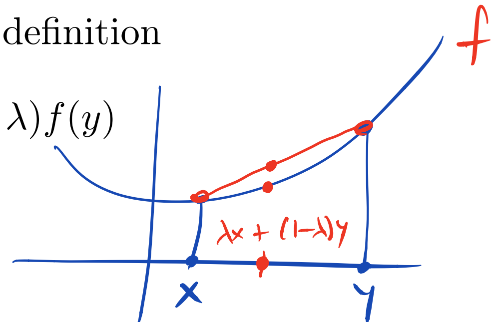
for all $x, y \in \mathbb{R}^n$.



Other equivalent definitions of convexity include positive definiteness of the Hessian matrix $\nabla^2 f(x)$ for all x, and the convexity along lines definition

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y)$$

for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$.



Convergence to a minimizer

Theorem 3. *Assume f is convex, ∇f is L-Lipschitz, and take $\alpha \leq \frac{1}{L}$. Then for any integer $t \geq 1$ we have*

$$f(x_t) - f_* \leq \frac{\|x_0 - x_*\|^2}{2\alpha t}, \tag{5}$$

$$= O\left(\frac{1}{t}\right)$$

where x_* is any minimizer of f.

Remark 4. Theorem 3 shows that the *values* $f(x_k)$ of gradient descent converge to the optimal value f_* at a rate of $O\left(\frac{1}{t}\right)$ when f is convex. This is an *extremely slow* convergence rate, known as *sublinear*. To get with $\varepsilon > 0$ of the optimal value requires $O\left(\varepsilon^{-1}\right)$ iterations. So if you want 10^{-6} accuracy you need 10^6 iterations.

$$f(x_t) - f_* \leq \frac{c}{t} = \varepsilon = 10^{-6}$$

Proof: Start with (for $\alpha \le \frac{1}{L}$).

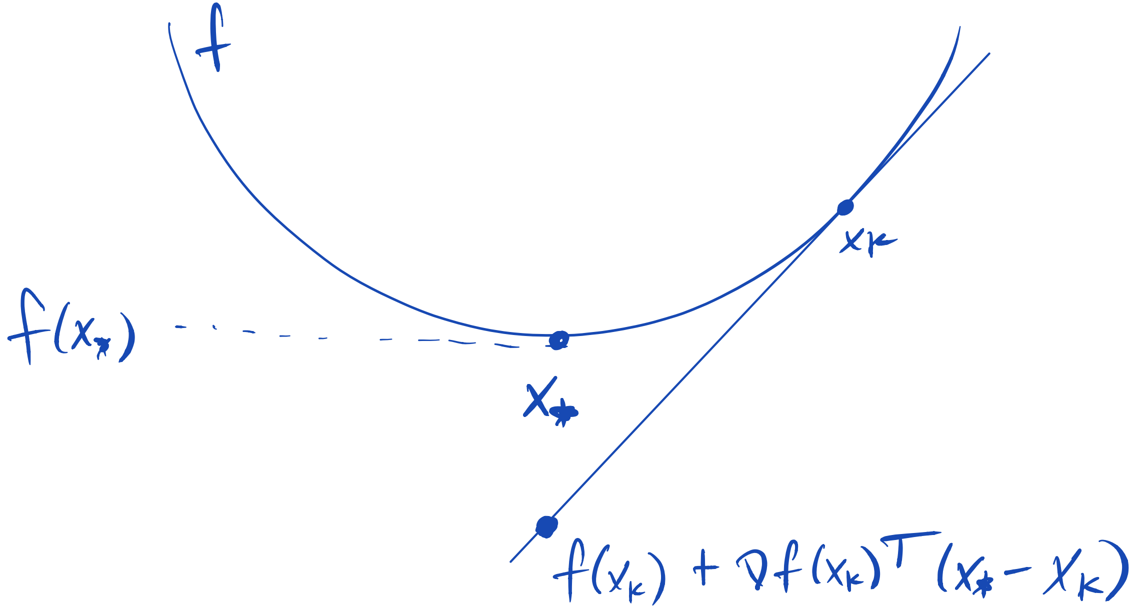
$$(*) \quad f(x_{k+1}) \le f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|^2.$$

Let $x_* \in \mathbb{R}^n$ be a minimizer of f, so

$$f_* = f(x_*).$$

Since f is convex ($y = x_*$, $x = x_k$)

$$f(x_*) \ge f(x_k) + \nabla f(x_k)^T (x_* - x_k)$$



Rearrange

$$f(x_k) \le f(x_*) + Df(x_k)^T (x_k - x_*)$$

Plug this into (*) to obtain

$$f(x_{k+1}) \le f(x_k) - \frac{\alpha}{2}\|\nabla f(x_k)\|^2$$

$$\le f(x_*) + \nabla f(x_k)^T(x_k - x_*) - \frac{\alpha}{2}\|\nabla f(x_k)\|^2$$

$$f(x_{k+1}) - f_* \le \nabla f(x_k)^T(x_k - x_*) - \frac{\alpha}{2}\|\nabla f(x_k)\|^2$$

$$= \frac{1}{2\alpha}\left(2\alpha\nabla f(x_k)^T(x_k - x_*) - \alpha^2\|\nabla f(x_k)\|^2\right)$$

$$\|x-y\|^2 = \|x\|^2 - 2x^Ty + \|y\|^2$$

$$= \frac{1}{2\alpha}\left(-\|\underbrace{x_k - x_* - \alpha\nabla f(x_k)}_{= x_{k+1} - x_*}\|^2 + \|x_k - x_*\|^2\right)$$

$$= \frac{1}{2\alpha}\left(\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\right).$$

Sum both sides

$$\sum_{k=0}^{t-1}\left(f(x_{k+1}) - f_*\right) \leq \frac{1}{2\alpha}\sum_{k=0}^{t-1}\left(\|x_k - x_*\|^2 - \|x_{k+1} - x_*\|^2\right)$$

$$= \frac{1}{2\alpha}\left(\|x_0 - x_*\|^2 - \|x_t - x_*\|^2\right)$$

$$\leq \frac{\|x_0 - x_*\|^2}{2\alpha}. \quad (**)$$

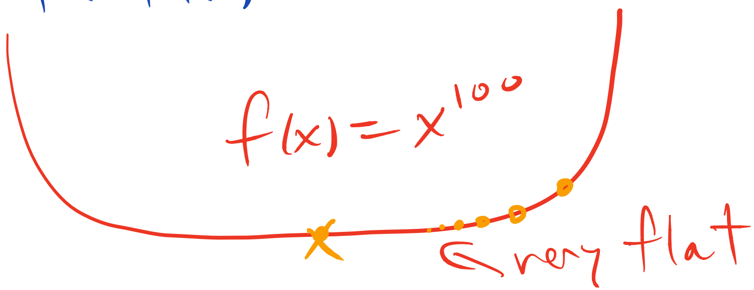
By $(*)$ $f(x_k)$ is decreasing and so

$$\sum_{k=0}^{t-1} \left(f(x_{k+1}) - f_* \right) \geq t \cdot \left(f(x_t) - f_* \right).$$

Plug into (**) to get

$$t \left(f(x_t) - f_* \right) \leq \frac{\|x_0 - x_*\|^2}{2\alpha} \qquad \blacksquare$$

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



Slow progress in flat regions.

Linear convergence

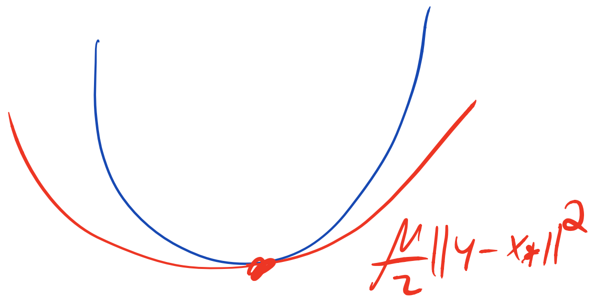
To obtain a better convergence rate, we need to make an additional assumption about how flat f can be at minima. We say that f is *μ-strongly convex* if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|x - y\|^2 \tag{6}$$

for all $x, y \in \mathbb{R}^n$.

Note: If we take $x = x_*$ then $\nabla f(x_*) = 0$ and we get

$$f(y) \geq f_* + \frac{\mu}{2}\|y - x_*\|^2. \tag{7}$$



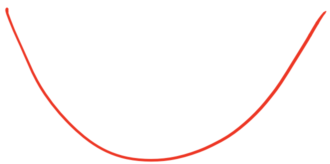
Polyak-Lojasiewicz (PL) inequality

If f is μ-strongly convex, then f satisfies the PL inequality

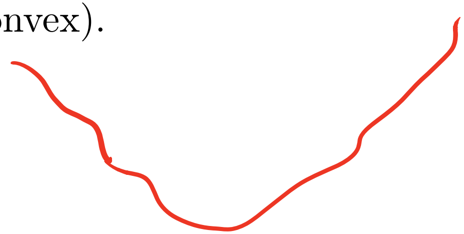
$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f_*) \tag{8}$$

for all $x \in \mathbb{R}^n$.

Remark 5. The PL inequality is weaker than strong convexity, and even nonconvex functions can satisfy it (as an exercise, show that $f(x) = x^2 + 3\sin^2(x)$ satisfies the PL inequality (8) with $\mu = \frac{1}{32}$, but f is not convex).



x^2



$f(x)$

Proof of PL-inequality: If f is

μ-strongly convex then for all $x, y \in \mathbb{R}^n$

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

Minimize both sides over $y \in \mathbb{R}^n$

$$f_* = \min_{y \in \mathbb{R}^n} f(y) \geq f(x) + \min_{y \in \mathbb{R}^n} \left\{ \nabla f(x)^T (y-x) + \frac{\mu}{2} \|x-y\|^2 \right\}$$

Take ∇ in y set $= 0$

$$0 = \nabla f(x) + \mu (y - x)$$

$$y - x = -\frac{1}{\mu} \nabla f(x)$$

$$f_* \geq f(x) + \nabla f(x)^T \left(-\frac{1}{\mu} \nabla f(x)\right) + \frac{\mu}{2} \left\| -\frac{1}{\mu} \nabla f(x) \right\|^2$$

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

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

$$\boxed{\frac{1}{2} \|\nabla f(x)\|^2 \geq \mu \left(f(x) - f_*\right)}$$

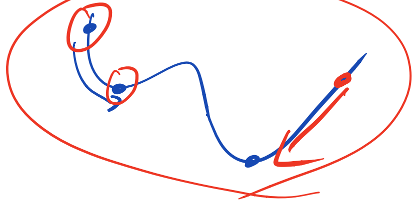
PL - inequality

General fact: If $f(x) \geq g(x)$ for all x then $\min\limits_x f(x) \geq \min\limits_x g(x)$

Take x_* s.t. $\min\limits_x f(x) = f(x_*)$

$$\min_x f(x) = f(x_*) \geq g(x_*) \geq \min_x g(x)$$

Simulated annealing



Linear convergence

Theorem 6. *Assume f satisfies the PL inequality (8), ∇f is L-Lipschitz, and take $\alpha \leq \frac{1}{L}$. Then for any integer $t \geq 0$ we have*

$$f(x_t) - f_* \leq (1 - \alpha\mu)^t (f(x_0) - f^*). \tag{9}$$

Proof: Start with (#)

$$f(x_{k+1}) \leq f(x_k) - \frac{\alpha}{2} \|\nabla f(x_k)\|^2$$

and apply PL-inequality

$$f(x_{k+1}) \leq f(x_k) - \alpha\mu (f(x_k) - f_*)$$

sub. f_* both sides.

$$
\begin{aligned}
f(x_{k+1}) - f_* &\leq (1-\alpha\mu)\,(f(x_k) - f_*) \\
&\leq (1-\alpha\mu)^2 (f(x_{k-1}) - f_*) \\
&\;\;\vdots \\
&\leq (1-\alpha\mu)^{k+1} (f(x_0) - f_*) \qquad \blacksquare
\end{aligned}
$$

Convergence of minimizers

Remark 7. It is also natural to ask how quickly x_k is converging to x_*. For this, we require strong convexity. If f is μ-strongly convex then we have

$$\frac{\mu}{2}\|x_t - x_*\|^2 \leq f(x_t) - f_* \leq (1 - \alpha\mu)^t (f(x_0) - f^*).$$