Mathematics of Image and Data Analysis
Math 5467

Gradient Descent

Instructor: Jeff Calder
Email: jcalder@umn.edu

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

Announcements

- HW2 due Feb 25

Last time

- PageRank

Today

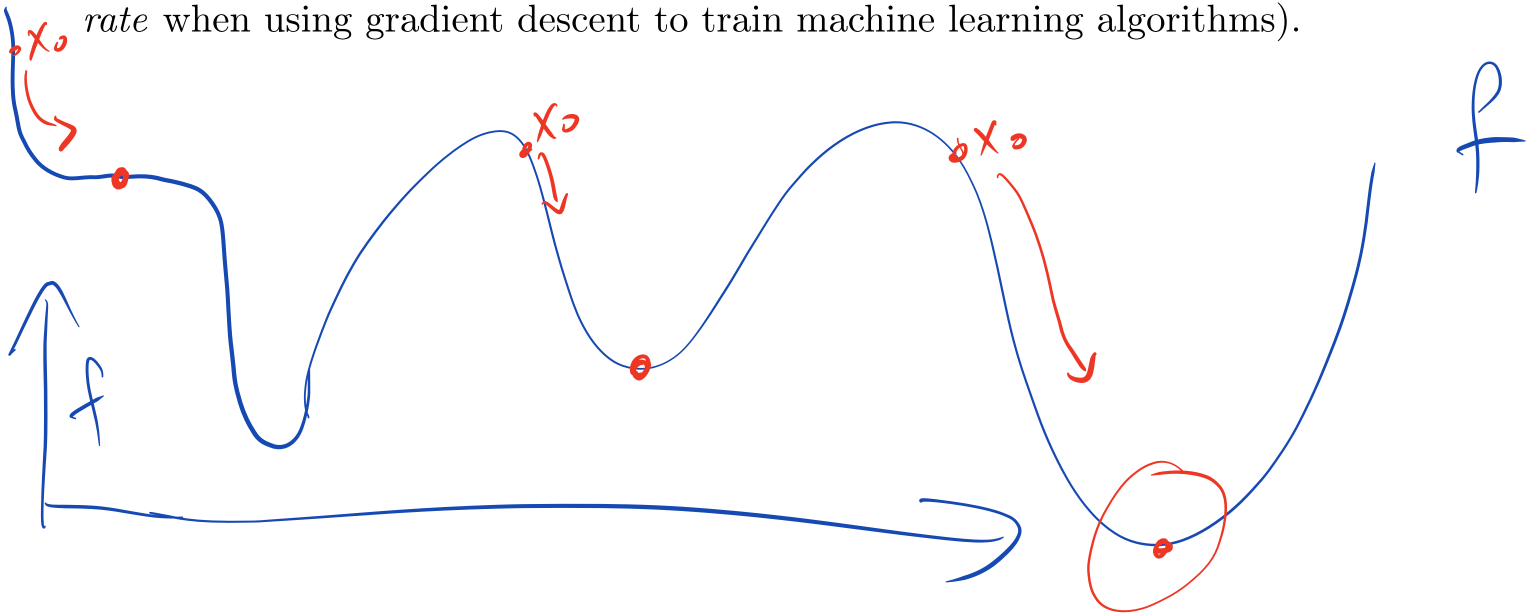
- Gradient Descent

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).



$\nabla(v^T x) = v$

Optimization interpretation

Exercise 1. Fix x_k and define

$$T(x) = \overbrace{f(x_k) + \nabla f(x_k)^T(x - x_k)}^{\text{hyper tangent plane at } x_k} + \underbrace{\frac{1}{2\alpha}\|x - x_k\|^2}_{\text{penalty}}. \tag{2}$$

If we define x_{k+1} as the minimizer of T, show that

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

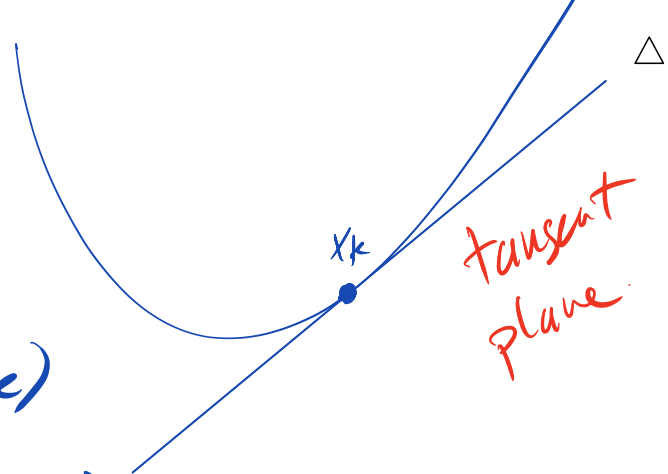
$\nabla T(x) = 0$

$\nabla T = \nabla f(x_k) + \frac{1}{\alpha}(x - x_k)$

$\nabla T = 0$

$\hookrightarrow \frac{1}{\alpha}(x - x_k) = -\nabla f(x_k)$

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



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

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n. \tag{3}$$

Theorem 2. *Assume ∇f is L-Lipschitz and that $\alpha \leq \frac{1}{L}$. Then for any integer $t \geq 1$ we have*

$$\min_{0\leq k\leq t} \|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f_*)}{\alpha t}. \tag{4}$$

Remark 3. 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.

$f: \mathbb{R}^n \to \mathbb{R}, \quad x, y \in \mathbb{R}^n$

Proof: Review Taylor expansion.

$$g(t) = f(x + t(y-x)), \quad t \in \mathbb{R}$$

$$g(1) = g(0) + \int_0^1 g'(t)\, dt \qquad (FTC)$$

$$= g(0) + \int_0^1 g'(0)\, dt + \int_0^1 g'(t) - g'(0)\, dt$$

$$g(1) = g(0) + g'(0) + R$$

$$R = \int_0^1 g'(t) - g'(0)\, dt$$

$$g'(t) = \frac{d}{dt} f(x + t(y-x))$$

$$= \nabla f(x+t(y-x))^T (y-x)$$

$$g'(0) = \nabla f(x)^T (y-x)$$

$$g(1) = g(0) + g'(0) + R$$

$$f(y) = f(x) + \nabla f(x)^T (y-x) + R$$

$$|R| = \left| \int_0^1 g'(t) - g'(0)\, dt \right|$$

$$\leq \int_0^1 |g'(t) - g'(0)|\, dt$$

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

$$(\ast) = \int_0^1 \left| \left[\nabla f(x+t(y-x)) - \nabla f(x)\right]^T (y-x) \right| dt$$

Aside Cauchy-Schwarz: $v^T w \le \underbrace{\|v\| \cdot \|w\|}_{=1}$

Proof: Assume $\|v\| = 1 = \|w\|$

$$0 \le \|v-w\|^2 = \|v\|^2 - 2v^Tw + \|w\|^2$$

$$= 2 - 2v^Tw = 2(1 - v^Tw)$$

$$\Rightarrow \quad 1 - v^Tw \ge 0 \Rightarrow v^Tw \le 1$$

$\blacksquare$

(*) Cauchy-Schwarz gives

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

since ∇f is L-Lipschitz

$$\le L\|x-y\| \int_0^1 \|t(y-x)\| dt$$

$$= L\|x-y\|^2 \int_0^1 t\, dt = \frac{L}{2}\|x-y\|^2$$

$$\boxed{f(y) = f(x) + \nabla f(x)^T(y-x) + R}$$

where $\boxed{|R| \leq \frac{L}{2}\|x-y\|^2}$

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

Theorem 2. *Assume ∇f is L-Lipschitz and that $\alpha \leq \frac{1}{L}$. Then for any integer $t \geq 1$ we have*

$$\min_{0\leq k\leq t} \|\nabla f(x_k)\|^2 \leq \frac{2(f(x_0) - f_*)}{\alpha t}. \tag{4}$$

Taylor expansion gives $x = x_k$, $y = x_{k+1}$

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

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

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

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

$$= f(x_k) - \alpha \underbrace{\left(1 - \frac{\alpha L}{2}\right)}_{\geq \frac{1}{2}} \|\nabla f(x_k)\|^2$$

$$\boxed{\text{Assume } \alpha \leq \frac{1}{L}}$$

$$1 - \frac{\alpha L}{2} \geq \frac{1}{2}$$

$$\frac{\alpha L}{2} \leq \frac{1}{2}$$

$$\alpha \leq \frac{1}{L}$$

Then

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

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

$$= \frac{2}{\alpha}\left(f(x_k) - f(x_{k+1})\right)$$

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

$$= \frac{2}{\alpha}\left(f(x_0) - f(x_{t+1})\right)$$

$$f(x_{t+1}) \geq f_* \qquad \leq \frac{2}{\alpha}\left(f(x_0) - f_*\right)$$

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

$$\boxed{\min_{0 \leq k \leq t} \|\nabla f(x_k)\|^2 \leq \frac{2}{\alpha} \frac{\left(f(x_0) - f_*\right)}{t+1}}$$

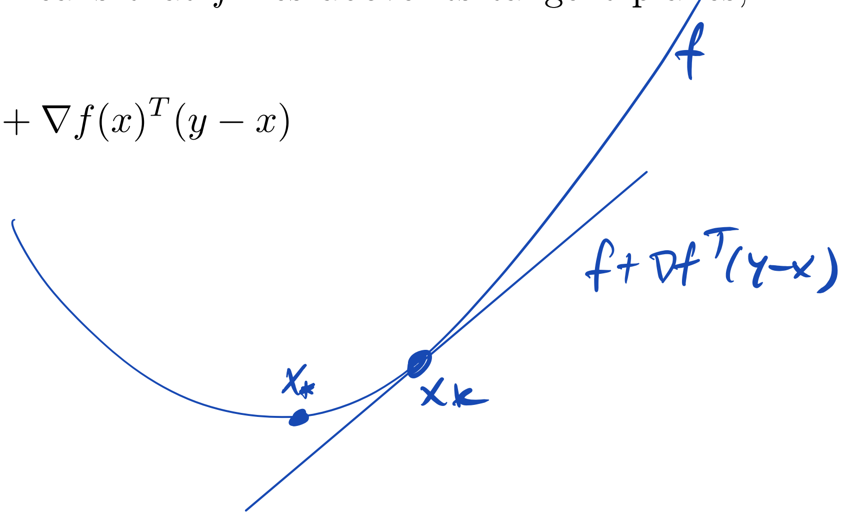
$\blacksquare$

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

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{5}$$

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

$$f(x_t) - f_* \leq \frac{\|x_0 - x_*\|^2}{2\alpha t}, \tag{6}$$

where x_* is any minimizer of f.

Remark 5. Theorem 4 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.

Proof: By convexity $\quad (f(x_*) = f_*)$

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

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

From prev. proof $\left(\alpha \leq \frac{1}{L}\right)$

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

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

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

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

$$2x^T y - \|y\|^2 = \|x\|^2 - \|x-y\|^2$$

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

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

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

$$= \frac{1}{2\alpha}\left(\|x_0 - x_*\|^2 - \cancel{\|x_t - x_*\|^2}\right)$$
$$\ge 0$$

$$\le \frac{\|x_0 - x_*\|^2}{2\alpha}$$

Note $\displaystyle\sum_{k=0}^{t-1} \left(f(x_{k+1}) - f_*\right) \ge t\left(f(x_t) - f_*\right)$

Hence $\displaystyle f(x_t) - f_* \le \frac{\|x_0 - x_*\|^2}{2\alpha t}.$ ▨

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

If ∇f is L-Lipschitz then

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

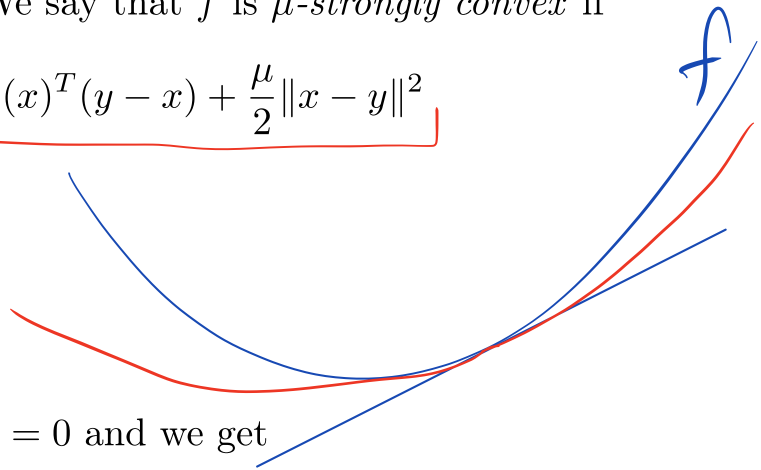
$$\Longrightarrow \mu \leq L$$

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

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|x - y\|^2 \tag{7}$$

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



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

$$f(y) \geq f_* + \frac{\mu}{2}\|y - x_*\|^2. \tag{8}$$

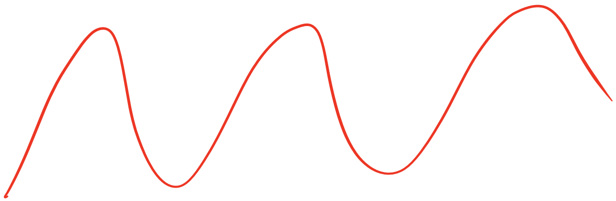
Polyak-Lojasiewicz (PL) inequality

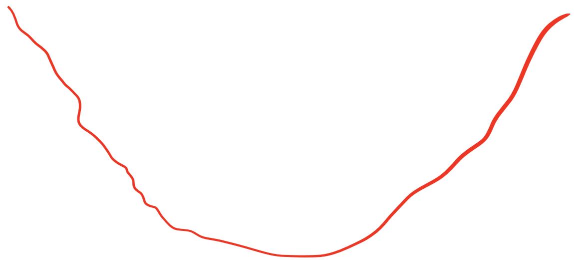
If f is μ-strongly convex, then f satisfies the PL inequality

$$\boxed{\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f_*)} \tag{9}$$

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

Remark 6. 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 (9) with $\mu = \frac{1}{32}$, but f is not convex).





Strong convexity

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

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

$$f_* = \min f \geq f(x) + \min_{y \in \mathbb{R}^n} \underbrace{\left\{ \nabla f(x)^T (y-x) + \frac{\mu}{2} \|x-y\|^2 \right\}}_{\nabla = 0}$$

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

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

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

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

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

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

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

PL-inequality.

$$\mu \le L \longrightarrow \frac{1}{L} \le \frac{1}{\mu}$$

$\Longrightarrow$ $\alpha \leq \frac{1}{\mu}$ $\Longrightarrow$ $\alpha\mu \leq 1$

Linear convergence

Theorem 7. *Assume f satisfies the PL inequality (9), ∇f is L-Lipschitz, and take $\alpha \leq \frac{1}{L}$. Then for any integer $t \geq 0$ we have*

$$f(x_t) - f_* \leq (1 - \alpha\mu)^t (f(x_0) - f^*). \tag{10}$$

Proof: From previous proof, for $\alpha \leq \frac{1}{L}$

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

$$\leq -\alpha\mu (f(x_k) - f_*)$$

PL-inequality

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

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

$$= f(x_k) - f_* - \alpha\mu\left(f(x_k) - f_*\right)$$

$$= (1-\alpha\mu)\left(f(x_k) - f_*\right)$$

$$\le (1-\alpha\mu)^2\left(f(x_{k-1}) - f_*\right)$$

$$\le (1-\alpha\mu)^3\left(f(x_{k-2}) - f_*\right)$$

$$\dots$$

$$f(x_k) - f_* \le (1-\alpha\mu)^k\left(f(x_0) - f_*\right) \quad \blacksquare$$

Convergence of minimizers

Remark 8. 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

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

Strong convexity

Gradient Descent Notebook (.ipynb)