Gradient Descent and Stochastic Gradient Descent

by Paul Hand
Northeastern University

Outline:

Gradient Descent (GD)

Convergence of GD

Stochastic Gradient Descent (SGD)

Analysis of SGD

Optimization and machine learning

Data $\{(x_i, y_i)\}_{i=1 \cdots n}$

Consider a model $\hat{y}_\theta(x_i)$

$$\min_\theta \sum_{i=1}^n \ell(\hat{y}_\theta(x_i), y_i)$$

Optimization in general

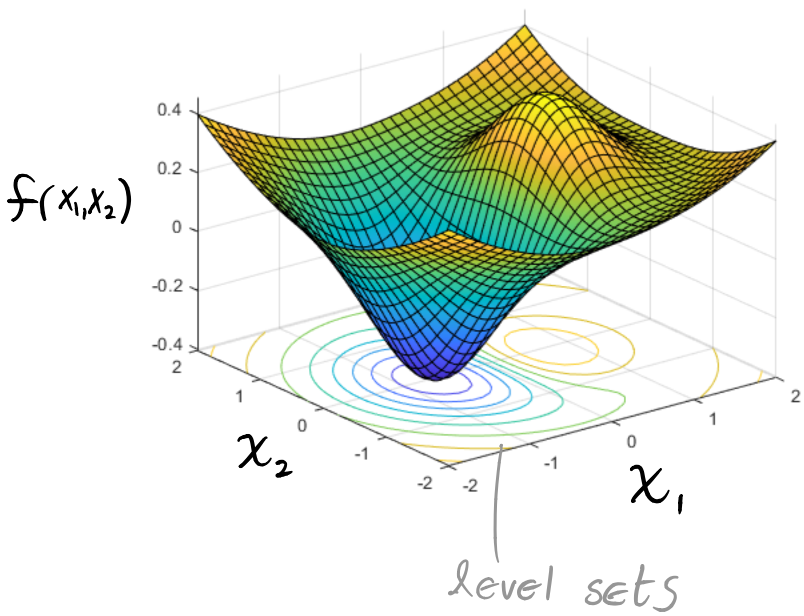
$$\min_x f(x)$$

Gradient descent: Take successive steps "downhill"

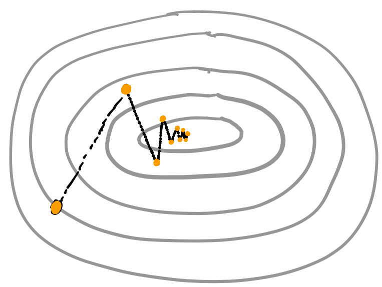
$$x^{i+1} = x^i - \alpha \nabla f(x^i)$$

α: step size, learning rate

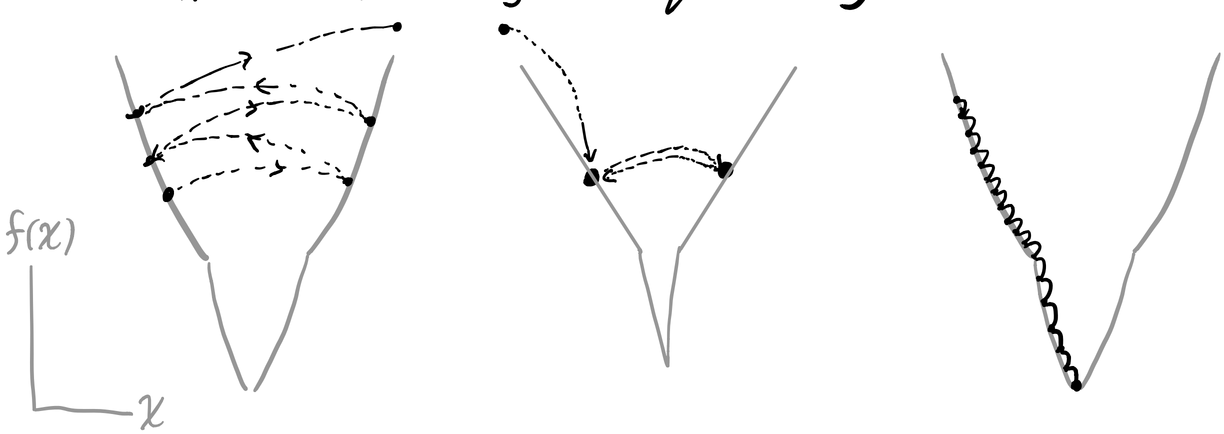
$-\nabla f$ points in direction of steepest descent



Depiction of gradient descent



How does learning rate qualitatively affect behavior?



too big l.r. can cause divergence

too big l.r. can miss a local well

too small l.r. can take a long time to converge

How fast does gradient descent converge?

$$\min_x f(x), \qquad x^{i+1} = x^i - \alpha \nabla f(x^i)$$

Suppose $x^i \to x^*$ as $i \to \infty$.

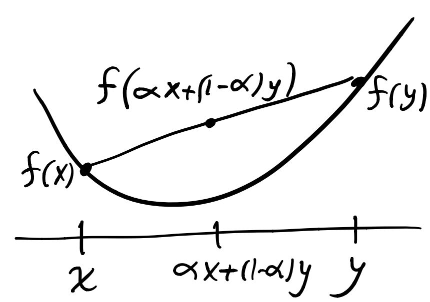
How long do you need to wait to get a certain accuracy ε?

Can gain understanding in some convex cases.

We say $f: \mathbb{R}^d \to \mathbb{R}$ is convex if

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

for all $0 \leq \alpha \leq 1$, x, y.



"always curves up"

f is convex if $D^2 f = Hf$ is positive semidefinite everywhere

Hessian matrix

$$D^2 f = Hf(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \vdots & & \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

H is positive definite if all eigenvalues are positive

H is positive semidefinite if all eigenvalues are nonnegative

Convergence of GD for quadratic functions

Let $f(x) = \frac{1}{2} x^t Q x - b^t x$

where $x \in \mathbb{R}^d$, $b \in \mathbb{R}^d$, $Q \in \mathbb{R}^{d \times d}$ is positive definite

Let $m = \lambda_{min}(Q)$, $M = \lambda_{max}(Q)$, $K = \frac{M}{m}$

condition number of Q

Consider GD w/ fixed step size α

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

Note: $x^* = Q^{-1} b$ is the unique global min of f

Theorem: If $\alpha = \frac{2}{M+m}$, then GD for $f(x) = \frac{1}{2}x^t Q x - b^t x$ satisfies

$$\|x^k - x^*\| \leq \left(\frac{1-1/\kappa}{1+1/\kappa}\right)^k \|x^0 - x^*\|$$

"first-order convergence"
Error decays exponentially
To get error ε, need $O(\log(\varepsilon^{-1}))$ iterations

Proof: Note $\nabla f(x) = Qx - b$.
The global minimizer solves $Qx^* = b \Rightarrow x^* = Q^{-1}b$

$$\begin{aligned}
x^{k+1} - x^* &= x^k - \alpha \nabla f(x^k) - x^* \\
&= x^k - \alpha(Qx^k - b) - x^* \\
&= x^k - \alpha(Qx^k - Qx^*) - x^* \\
&= (I - \alpha Q)(x^k - x^*)
\end{aligned}$$

So,

$$\|x^{k+1} - x^*\| \leq \underbrace{\|I - \alpha Q\|}_{\max(\alpha M - 1,\, 1 - \alpha m)} \|x^k - x^*\|$$

We choose $\alpha = \frac{2}{M+m}$.

So $\|I - \alpha Q\| = \frac{M-m}{M+m} = \frac{1 - 1/k}{1 + 1/k} < 1$

$$\Rightarrow \|x^{k+1} - x^*\| \le \left(\frac{1 - 1/k}{1 + 1/k}\right) \|x^k - x^*\|$$

$$\Rightarrow \|x^k - x^*\| \le \left(\frac{1 - 1/k}{1 + 1/k}\right)^k \|x^0 - x^*\| \quad \blacksquare$$

Interpretation:

If f doesnt curve up too much and doesnt curve up too little, then GD with fixed step size can exhibit first order convergence to the global minimizer

Convergence of GD for convex strongly smooth f

Defn: f is **M-Strongly smooth** if

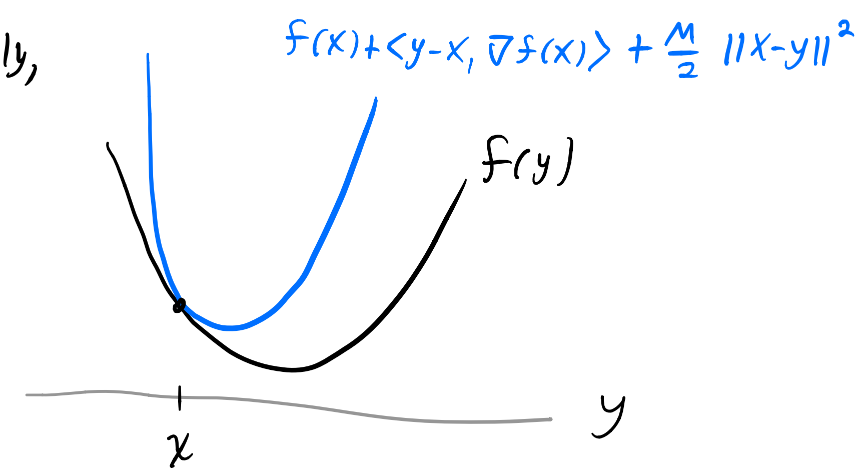
$$\forall\, x, y \qquad f(y) - f(x) \leq \langle y - x, \nabla f(x) \rangle + \frac{M}{2} \|y - x\|^2$$

or, equivalently

$$\|\nabla f(x) - \nabla f(y)\| \leq M \|x - y\|$$

∇f is M-Lipschitz

Visually,



"f doesn't curve up too much"

Theorem: Let f be convex and M-strongly smooth. If $\alpha \leq \frac{1}{M}$, then GD satisfies

$$f(x^i) - f(x^*) \leq \frac{1}{2i\alpha} \|x^0 - x^*\|^2$$

where x^* is a minimizer of f.

- Error decays slowly
- To get error ε from optimal value, need $O(\varepsilon^{-1})$ iterations

Proof:

$$\begin{aligned} f(x^{k+1}) - f(x^k) &\leq \langle x^{k+1} - x^k, \nabla f(x^k) \rangle + \frac{M}{2} \|x^{k+1} - x^k\|^2 \\ &= -\alpha \|\nabla f(x^k)\|^2 + \frac{M}{2}\alpha^2 \|\nabla f(x^k)\|^2 \\ &= -\alpha\left(1 - \frac{\alpha M}{2}\right) \|\nabla f(x^k)\|^2 \\ &\leq -\frac{\alpha}{2} \|\nabla f(x^k)\|^2 \end{aligned}$$

Note: $f(x^k) - f(x^*) \leq \langle x^k - x^*, \nabla f(x^k) \rangle$ by convexity

So,
$$
\begin{aligned}
f(x^{k+1}) &\leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&\leq f(x^*) + \langle x^k - x^*, \nabla f(x^k)\rangle - \frac{\alpha}{2}\|\nabla f(x^k)\|^2 \\
&= f(x^*) + \frac{1}{2\alpha}\left(\|x^k - x^*\|^2 - \|x^k - x^* - \alpha\nabla f(x^k)\|^2\right) \\
&= f(x^*) + \frac{1}{2\alpha}\left(\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2\right)
\end{aligned}
$$

So,
$$
\begin{aligned}
f(x^k) - f(x^*) &\leq \frac{1}{k}\sum_{j=1}^{k} f(x^j) - f(x^*) \qquad (\text{as } f(x^k) \text{ is decreasing in } k) \\
&\leq \frac{1}{2k\alpha}\left(\|x^0 - x^*\|^2 - \|x^k - x^*\|^2\right) \\
&\leq \frac{1}{2k\alpha}\|x^0 - x^*\|^2
\end{aligned}
$$

$\blacksquare$

Challenges of gradient descent in deep learning

$$\min_{\theta} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(\hat{y}_\theta(x_i), y_i)}_{f(\theta)}$$

$$\theta^{k+1} = \theta^k - \alpha \nabla f(\theta) = \theta^k - \alpha \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(\hat{y}_\theta(x_i), y_i)$$

To evaluate $\nabla f(\theta)$, one needs to loop through all data (batch gradient descent)

- expensive
- not possible in some contexts

Idea: use minibatches

Select a minibatch $B \subset \{1, 2, \dots, n\}$

$$\theta^{k+1} = \theta^k - \alpha \underbrace{\frac{1}{|B|}\sum_{i \in B} \nabla_\theta \ell(\hat{y}_\theta(x_i), y_i)}_{\text{use as approximation of } \nabla_\theta f(\theta)}$$

If the minibatch is chosen randomly, on average, the gradient of a minibatch is the full gradient

$\Rightarrow$ Stochastic gradient descent

Stochastic Gradient Descent

Want to solve $\min_x f(x)$

Instead of having access to $\nabla f(x)$, suppose only have $G(x)$ w/ $\mathbb{E}[G(x)] = \nabla f(x)$.

Write SGD as

$$x^{k+1} = x^k - \alpha_k G(x^k)$$

- on average, move in direction of steepest descent
- may move further from minimizer

Simple model: additive noise

$$G(x) = \nabla f(x) + w, \qquad w \sim N(0, \sigma^2 I)$$

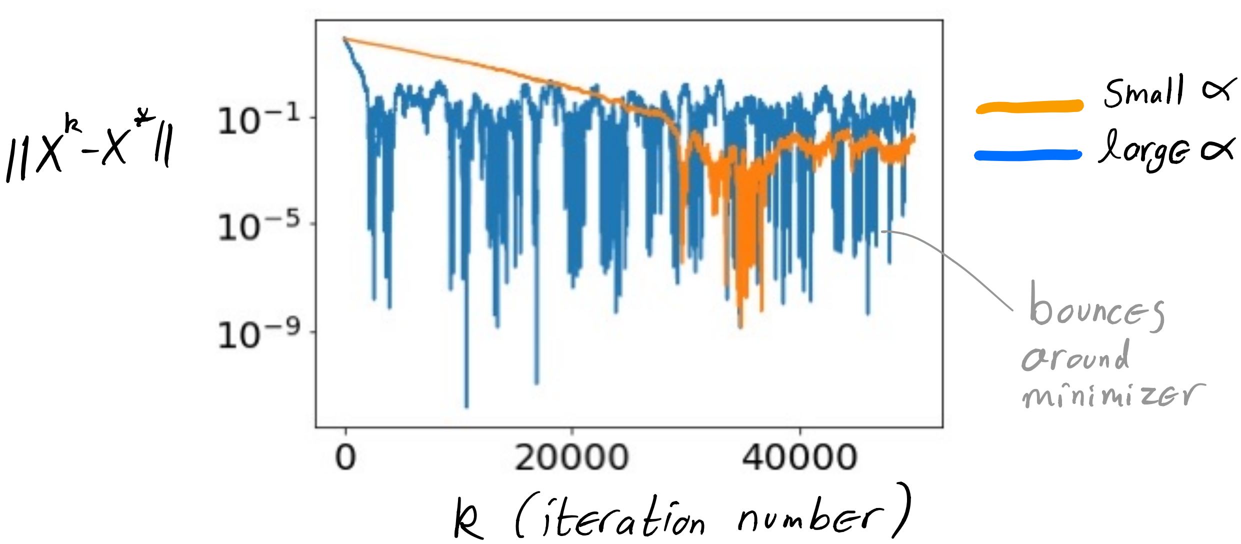
Use in ML: minibatches

$$f(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\hat{y}_\theta(x_i), y_i)$$

$$G(\theta) = \frac{1}{|B|} \sum_{i \in B} \nabla_\theta \ell(\hat{y}_\theta(x_i), y_i) \quad \text{for random subset } B$$

Qualitatively,

with fixed step size α, x^k will move close to x^* but will bounce around due to stochasticity



large $\alpha \Rightarrow$ fast initial convergence
large error

Small $\alpha \Rightarrow$ slow initial convergence
smaller error

Can formalize these observations w/ theory

Analysis of SGD

Consider a convex $f : \mathbb{R}^d \to \mathbb{R}$
Suppose $\mathbb{E}(G(x)) = \nabla f(x)$

We say the stochastic gradient is
(M, B)-bounded if

$$\mathbb{E}\, \|G(x)\|^2 \leq M^2 \|x - x_*\|^2 + B^2$$

where x^* is a minimizer of f.

Examples:

- If $G(X) = \nabla f(X)$ and f is M-strongly smooth,

  G is $(M, 0)$ bounded

- If $G(X) = \nabla f(X) + W$ w/ $W \sim N(0, \sigma^2 I)$

  $\mathbb{E}\|G(X)\|^2 = \|\nabla f(X)\|^2 + \mathbb{E}[\|W\|^2]$

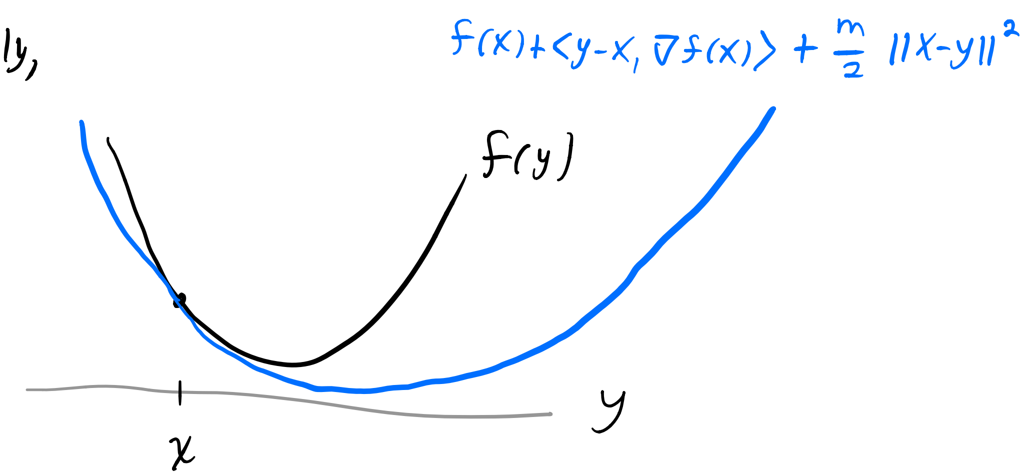
  If f is M-strongly smooth

  G is $(M, \sqrt{d}\,\sigma)$ bounded

Defn: f is m-strongly convex if

$$\forall x, y, \quad f(y) \geq f(x) + \langle y - x, \nabla f(x) \rangle + \frac{m}{2} \|y - x\|^2$$

Visually,



"f doesn't curve up too little"

Theorem:

If f is m-strongly convex and G is (M, B)-bounded, and $\alpha \in (0, \frac{m}{M^2})$

$$\mathbb{E}\, \|x^k - x^*\|^2 \le \underbrace{(1 - 2m\alpha + \alpha^2 M^2)^k \|x^0 - x^*\|^2}_{\text{looks like first order convergence}} + \underbrace{\frac{\alpha B^2}{2m - \alpha M^2}}_{\text{up to some error}}$$

Note: For constant α, do not expect convergence.

Smaller α brings us closer to x^* but with slower convergence rate initially

How to choose step sizes/learning rates?

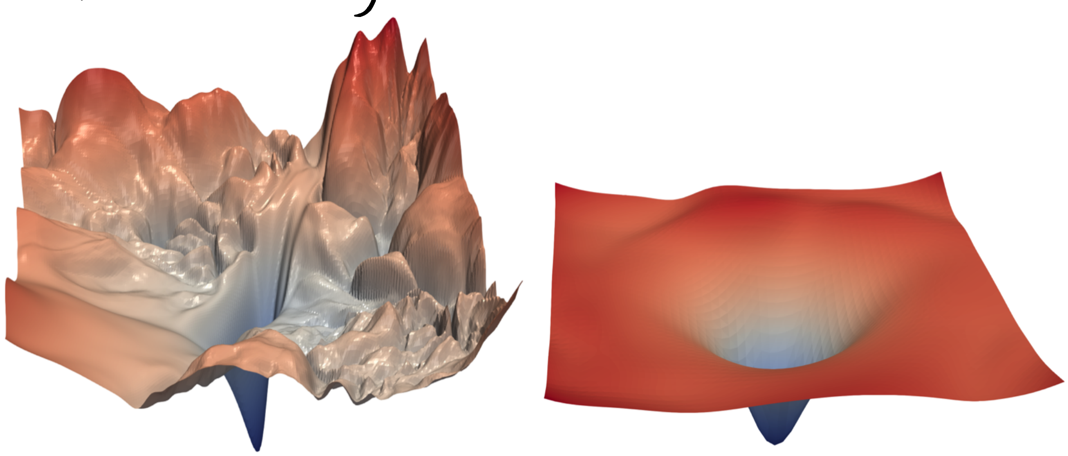
- Run at a large value for a while
- Shrink learning rate
- Repeat

- Have schedule of α_k decaying in k

In these cases can hope for convergence

Challenges w/ GD and SGD in Deep Learning

Nonconvexity and nonsmoothness



(a) without skip connections (b) with skip connections

Figure 1: The loss surfaces of ResNet-56 with/without skip connections. The proposed filter normalization scheme is used to enable comparisons of sharpness/flatness between the two figures.

(Li et al. 2018)

may be stuck in a local minimum, so may want to temporarily increase learning rate to get unstuck.

Summary:

- Too large learning rate can lead to divergence
- In convex case, to get convergence α should be small relative to curvature of f
- Too small learning rate can lead to slow convergence
- For convex quadratic functions, convergence of GD can be first order (fast)
- For more general convex functions, convergence can be slow
- SGD w/ fixed step size is not expected to converge
- SGD with decaying step sizes may converge