BGSM/CRM

AL&DNN

Gradient descent and stochastic approximation

S. Xambó

UPC & IMTech

14/10/2021

Abstract

A study of the Stochastic Gradient Descend (SGD) and its role in Deep Learning.

Introduced in [1], stochastic approximation has ever since been the focus of attention by many researchers.

Here are some of the sources appeared in the last decade that you may find useful for the study of today's topic:

[2], [3], [4], [5], [6], [7], [8], [9], [10], [11].

As a main reference you may consider [3, Ch. 14]

Index

Background notions

Directional derivatives, differentials and gradients
Epigraph of a function

- $\mathcal{X}$ is open subset of $\mathbf{R}^n$ and $f : \mathcal{X} \to \mathbf{R}$ is a differentiable function. The *directional derivative* of f at x in the direction v is
  $D_v f(x) = \frac{d}{dt} f(x + tv)|_{t=0}$.

- Since $f(x + tv) = f(x) + t(d_x f)(v) + O(t^2)$, by definition of the differential, we see that $D_v f(x) = (d_x f)(v)$.

- If $e_1, \cdots, e_n$ is the standard basis of $\mathbf{R}^n$, then
  $(d_x f)(e_i) = D_{e_i} f(x) = \partial f(x)/\partial x_i = \partial_i f(x)$.

- It follows that $(d_x f)(v) = \sum_{i=1}^n v_i \partial_i f(x) = v \cdot \nabla f(x)$, where
  $\nabla f(x) = (\partial_1 f(x), \cdots, \partial_n f(x))$.

- Therefore $D_v f(x) = v \cdot \nabla f(x)$. This implies that $\nabla f(x)$ is the direction of the greatest growth rate of f at x. Hence $-\nabla f(x)$ is the direction of steepest descent.

- $\nabla f(x)$ is orthogonal to the level sets $\mathcal{X}_\lambda = \{x \in \mathcal{X} \mid f(x) = \lambda\}$: if v is tangent to $\mathcal{X}_\lambda$, then $v \cdot \nabla f(x) = D_v f(x) = d_x f(v) = 0$.

$\mathcal{X}$ a subset of $\mathbf{R}^n$ and $f : \mathcal{X} \to \mathbf{R}$ a function.

The *epigraph* of a f, denoted $\mathrm{Epi}(f)$, is the subset of $\mathcal{X} \times \mathbf{R}$ whose points (x, t) satisfy $t \geqslant f(x)$.

Lemma. If $\mathcal{X}$ and f are convex, then $\mathrm{Epi}(f)$ is convex.

Proof. Let $(x, t), (x', t') \in \mathrm{Epi}(f)$. Choose any $\lambda \in (0, 1)$. We want to see that
$\lambda(x, t) + (1 - \lambda)(x', t') = (\lambda x + (1 - \lambda)x', \lambda t + (1 - \lambda)t') \in \mathrm{Epi}(f)$.
Since $\lambda x + (1 - \lambda)x' \in \mathcal{X}$, because $\mathcal{X}$ is convex, we can write:

$$\begin{aligned} f(\lambda x + (1 - \lambda)x') &\leqslant \lambda f(x) + (1 - \lambda)f(x') \quad \text{(as } f \text{ is convex)} \\ &\leqslant \lambda t + (1 - \lambda)t' \quad \text{(definition of epigraph).} \end{aligned}$$ □

Subgradients

... and convexity

Remarks

Examples

... and Lipschitzness

$\mathcal{X}$ a subset of $\mathbf{R}^n$ and $f : \mathcal{X} \to \mathbf{R}$.

- A vector $s \in \mathbf{R}^n$ is a *subgradient* of f at x if for any $x' \in \mathcal{X}$
  $f(x') \geqslant f(x) + s \cdot (x' - x)$, or $f(x) \leqslant f(x') + s \cdot (x - x')$.

Mnemonics:

$$f(x) - f(x') \leqslant s_x \cdot (x - x'), \quad f(x) - f(x') \geqslant s_{x'} \cdot (x - x')$$

- The *set* of subgradients of f at x is denoted $\partial f(x)$.

Theorem. Assume $\mathcal{X}$ is convex.
(a) If $\partial f(x) \neq \emptyset$ for all $x \in \mathcal{X}$, then f is convex.
(b) Conversely, if f is convex then $\partial f(x) \neq \emptyset$ for any $x \in \mathcal{X}^\circ$.
(c) If f is convex and differentiable at x, then $\nabla f(x) \in \partial f(x)$.

Proof. (a) Let $x, x' \in \mathcal{X}$ and $\lambda \in (0,1)$. We want to prove that $f(\lambda x + (1-\lambda)x') \leqslant \lambda f(x) + (1-\lambda)f(x')$.

Let $x_\lambda = (1-\lambda)x + \lambda x'$ and $s \in \partial f(x_\lambda)$. Then

$$f(x) \geqslant f(x_\lambda) + s\cdot(x - x_\lambda) = f(x_\lambda) + (1-\lambda)s\cdot(x-x'),$$
$$f(x') \geqslant f(x_\lambda) + s\cdot(x' - x_\lambda) = f(x_\lambda) + \lambda s\cdot(x'-x) \Rightarrow$$

$$\lambda f(x) + (1-\lambda)f(x') \geqslant f(x_\lambda).$$

(b) Let $x \in \mathcal{X}$. Then $(x, f(x)) \in \partial \mathrm{Epi}(f)$. Since $\mathrm{Epi}(f)$ is convex, by the separation hyperplane theorem there exists $(u, a) \in \mathbf{R}^n \times \mathbf{R}$, $(u, a) \neq (0, 0)$, such that

$(*)$ $\quad u \cdot x + af(x) \geqslant u \cdot x' + at'$ for all $(x', t') \in \mathrm{Epi}(f)$.

Since t' can be as large as we wish, we infer that $a \leqslant 0$.

Now let $x \in \mathcal{X}^\circ$. For a sufficiently small $\epsilon > 0$, $x' = x + \epsilon u \in \mathcal{X}$ and hence $u \cdot x + af(x) \geqslant u \cdot x + \epsilon u \cdot u + at'$, or $af(x) \geqslant \epsilon u \cdot u + at'$. This implies that $a < 0$: if $a = 0$, then $\epsilon u \cdot u \leqslant 0$, which is not possible because $(u, a) \neq (0, 0)$.

Set $t' = f(x')$ in the inequality $(*)$. Rearranging,

$$a(f(x') - f(x)) \leqslant u \cdot (x - x'), \text{ or } f(x') - f(x) \geqslant \frac{1}{-a} u \cdot (x' - x),$$

which shows that $s = \frac{1}{-a} u$ is a subgradient of f at x.

(c) If f is convex and differentiable at x, we know that

$$f(x') \geqslant f(x) + (x' - x) \cdot \nabla f(x).$$

But this just says that $\nabla f(x)$ is a subgradient of f at x. □

- It may be instructive to prove statement (c) in the present context. Rewrite the convexity condition of f,

$$f((1-\lambda)x+\lambda x') \leqslant (1-\lambda)f(x)+\lambda f(x')$$

in this form:

$$f(x') \geqslant \frac{f(x+\lambda(x'-x))-f(x)+\lambda f(x)}{\lambda}$$

$$= f(x)+\frac{f\left(x+\lambda(x'-x)\right)-f(x)}{\lambda}.$$

Now letting $\lambda \to 0$ in the fraction, we get $(x'-x)\cdot\nabla f(x)$, and this ends the proof. □

- In the statement (b), the condition $x \in \mathcal{X}^\circ$ can be replaced by $x \in \mathcal{X}^{\mathrm{ri}}$, the interior of $\mathcal{X}$ relative to its affine span $[\mathcal{X}]$.

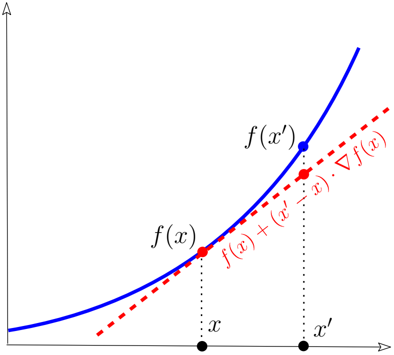
- $\nabla f(x)$ provides only local information about f around x, whereas $s \in \partial f(x)$ gives a linear function that is a (global) lower bound of f.

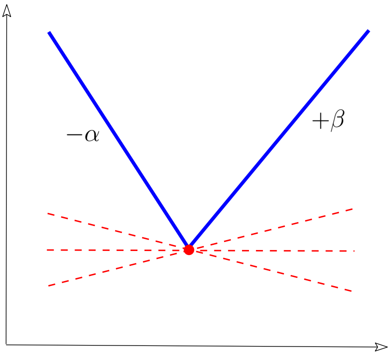
- A local minimum x of a convex function f is a global minimum (equivalent to $0 \in \partial f(x)$): For any x' and sufficiently small ϵ, $f(x) \leqslant f((1-\epsilon)x + \epsilon x') \leqslant (1-\epsilon)f(x) + \epsilon f(x') \Rightarrow f(x) \leqslant f(x')$.

Theorem. Let $\mathcal{X}$ be convex and closed, and $f : \mathcal{X} \to \mathbf{R}$ convex. Then $\bar{x} \in \operatorname{argmin}_{x\in\mathcal{X}} f(x)$ if and only if $\nabla f(\bar{x}) = 0$.

Proof. Assume $\bar{x} \in \mathcal{X}$ satisfies $f(\bar{x}) \leqslant f(x)$ for all $x \in \mathcal{X}$. Then in particular $h(t) = f(\bar{x} + t(x - \bar{x}))$ has a minimum at $t = 0$. So $\frac{dh(t)}{dt}|_{t=0} = 0$. But since this derivative is equal to $D_{x-\bar{x}} f(\bar{x}) = (x - \bar{x}) \cdot \nabla f(\bar{x})$, we have that $\nabla f(\bar{x})$ is orthogonal to all vectors of the form $x - \bar{x}$, $x \in \mathcal{X}$. But $\nabla f(\bar{x})$ belongs to the linear span of these vectors, and hence must vanish.

And if $\nabla f(\bar{x}) = 0$, then 0 is a subgradient of f at $\bar{x}$ and therefore $f(x) \geqslant f(\bar{x}) + 0 \cdot (x - \bar{x}) = f(\bar{x})$.





If $f(x)$ is differentiable at x, then $\nabla f(x)$ is the unique subgradient of f at x, and this gives the tangent at $(x, f(x))$ to the graph of f. The image on the left illustrates this. The function depicted on the right has constant slope $-\alpha$ $(+\beta)$ to the left (right) of x_0, so these are the only subgradients to the left (right) of x_0. At the point x_0, the subgradients are the points in the interval $[-\alpha, +\beta]$.

Example. Let $f_j(x)$, $j \in [m]$, be convex differentiable functions defined on a convex set $\mathcal{X}$.

Set $f(x) = \max_j f_j(x)$.

If for a given $x \in \mathcal{X}$ we have $f(x) = f_k(x)$, $k \in [m]$, then $\nabla f_k(x) \in \partial f(x)$.

Note that the function $f(x)$ is convex: if $x, x' \in \mathcal{X}$, and $\lambda \in (0,1)$, for any $j \in [m]$ we have

$f_j(\lambda x + (1-\lambda)x') \leqslant \lambda f_j(x) + (1-\lambda) f_j(x') \leqslant \lambda f(x) + (1-\lambda) f(x')$,

and hence $f(\lambda x + (1-\lambda)x') \leqslant \lambda f(x) + (1-\lambda) f(x')$.

Now we have: $f_k(x') \geqslant f_k(x) + (x'-x)\cdot\nabla f_k(x)$, as f_k is convex.

Since $f(x') \geqslant f_k(x')$ and $f_k(x) = f(x) \Rightarrow$

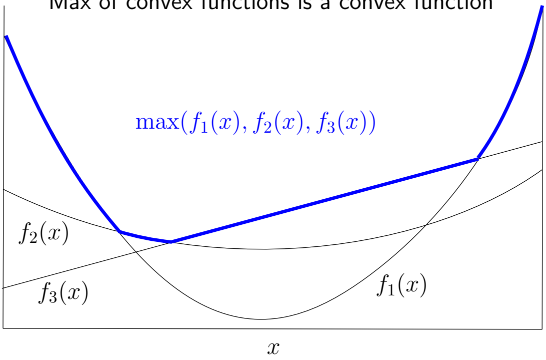
$$f(x') \geqslant f(x) + (x'-x)\cdot\nabla f_k(x). \qquad \square$$

A special case of the previous example is the *hinge loss*

$$f(x) = \max(0, 1 - y(x \cdot \boldsymbol{\xi}))$$

at a data point $\boldsymbol{\xi}$ with label $y \in \{\pm 1\}$. If $1 - y(x \cdot \boldsymbol{\xi}) < 0$, then 0 is a subgradient. Otherwise, it is $\nabla_x(1 - y(x \cdot \boldsymbol{\xi})) = -y\boldsymbol{\xi}$.

Max of convex functions is a convex function



Lemma. Let $\mathcal{X}$ be *open* and *convex* and let $f : \mathcal{X} \to \mathbf{R}$ be *convex*. Then f is ρ-Lipschitz over $\mathcal{X}$ if and only if $\|s\| \leqslant \rho$ for any $x \in \mathcal{X}$ and any $s \in \partial f(x)$.

Proof. ($\Leftarrow$) Assume that for all $x \in \mathcal{X}$ and $s \in \partial f(x)$ we have $\|s\| \leqslant \rho$. Then, for any $x' \in \mathcal{X}$, $f(x) - f(x') \leqslant s \cdot (x - x')$, by definition of subgradient, and

$$s \cdot (x - x') \leqslant \|s\|\|x - x'\| \leqslant \rho\|x - x'\| \text{ (by Cauchy-Schwartz).}$$

So $f(x) - f(x') \leqslant \rho\|x - x'\|$. Analogously, with $s' \in \partial f(x')$,

$$f(x') - f(x) \leqslant s' \cdot (x' - x) \leqslant \|s'\|\|x' - x\| \leqslant \rho\|x' - x\|.$$

In sum, $|f(x') - f(x)| \leqslant \rho\|x' - x\|$ and f is ρ-Lipschitz.

($\Rightarrow$) Assume f is ρ-Lipschitz and pick $x \in \mathcal{X}$ and $s \in \partial f(x)$.

Since $\mathcal{X}$ is open, there exists $\epsilon > 0$ such that

$$x' = x + \epsilon s/\|s\| \in \mathcal{X}.$$

Therefore

$$(x' - x) \cdot s = \epsilon\|s\| \text{ and } \|x' - x\| = \epsilon.$$

By the definition of subgradient,

$$f(x') - f(x) \geqslant s \cdot (x' - x) = \epsilon\|s\|.$$

On the other hand, by ρ-Lipschitzness,

$$\rho\epsilon = \rho\|x' - x\| \geqslant f(x') - f(x).$$

So

$$\epsilon\|s\| \leqslant f(x') - f(x) \leqslant \rho\epsilon,$$

and hence $\|s\| \leqslant \rho$. $\square$

Corollary. If f is differentiable and ρ-Lipschitz, then $\|\nabla f(x)\| \leqslant \rho$ for all x.

Proof. Its a direct consequence of the lemma on page 16 and the fact that the gradient $\nabla f(x)$ is a subgradient. $\square$

Gradient descent (GD)

Basic algorithms
Convergence results

Inputs

$f : \mathbf{R}^n \to \mathbf{R}$, $\eta \in \mathbf{R}_{++}$ (*learning rate*),
$x^0 \in \mathbf{R}^n$ (starting point), r (number of steps)

Procedure

Do r times:

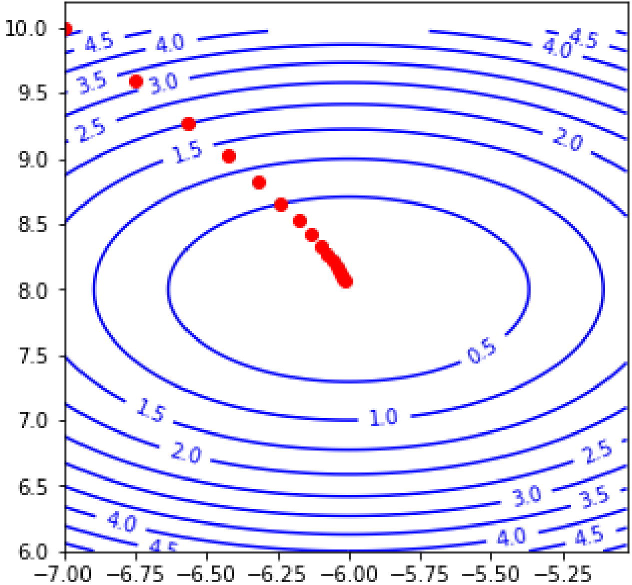
$$x^k = x^{k-1} - \eta \nabla f(x^{k-1})$$

Naif output: x^r.

Smart output: $\hat{x} = \frac{1}{r} \sum_{k \in [r]} x^k$.

Example (cf. [3, Fig. 14.1]). $f(x, y) = 1.25(x+6)^2 + (y-8)^2$,
$\nabla_{x,y} f = \big(2.5(x+6), 2(y-8)\big)$.
With $\eta = 0.1$, $x^0 = (-7, 10)$, and $r = 15$, the sequence $x^0, x^1, \cdots, x^r$ is depicted in the image on next page.

The blue lines represent level sets of $f(x, y)$.



10.0
9.5
9.0
8.5
8.0
7.5
7.0
6.5
6.0
−7.00
−6.75
−6.50
−6.25
−6.00
−5.75
−5.50
−5.25
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5

```
eta = 0.1
def f(x,y): return 1.25*(x + 6)**2 + (y-8)**2
def Gf(x,y): return [2.5*(x+6),2*(y-8)]

a = -7; b = 10
A=[a]; B=[b]
N = 15

for _ in range(1,N+1):
    ga,gb = Gf(a,b)
    a,b = (a-eta*ga, b-eta*gb)
    A += [a]; B += [b]

plt.plot(A,B,'o',color='r')
```

1. **Input**: Initial value $x = x^0$
2. **while not converged**:
3. $x = x - \eta \nabla f(x)$
4. **convergence check**
5. [**update** η]
6. **return** x.

Lemma

(a) Fix a positive integer r, a positive real number η, a vector $\bar{x} \in \mathbf{R}^n$, and a sequence $v^1, \ldots, v^r \in \mathbf{R}^n$. Let $x^1 = 0$ and define

$$x^{k+1} = x^k - \eta v^k \text{ for } k \in [r].$$

Then we have the inequality

$$\sum_{k\in[r]} \langle x^k - \bar{x}, v^k \rangle \leqslant \frac{1}{2\eta}\|\bar{x}\|^2 + \frac{\eta}{2}\sum_{k\in[r]} \|v^k\|^2. \tag{1}$$

(b) Fix $B, \rho \in \mathbf{R}_{++}$ such that $\|v^k\| \leqslant \rho$ and $\|\bar{x}\| \leqslant B$. Let $\eta = B/\rho\sqrt{r}$. Then

$$\frac{1}{r}\sum_{k\in[r]} \langle x^k - \bar{x}, v^k \rangle \leqslant B\rho/\sqrt{r}.$$

Proof. Using the identity $x \cdot x' = \frac{1}{2}(-\|x - x'\|^2 + \|x\| + \|x'\|^2)$ $(x, x' \in \mathbf{R}^n)$, we have:

$$\begin{aligned}\langle x^k - \bar{x}, v^k\rangle &= \frac{1}{\eta}\langle x^k - \bar{x}, \eta v^k\rangle \\ &= \frac{1}{2\eta}\left(-\|x^k - \bar{x} - \eta v^k\|^2 + \|x^k - \bar{x}\|^2 + \eta^2\|v^k\|^2\right) \\ &= \frac{1}{2\eta}\left(-\|x^{k+1} - \bar{x}\|^2 + \|x^k - \bar{x}\|^2\right) + \frac{\eta}{2}\|v^k\|^2.\end{aligned}$$

Adding up for $k \in [r]$, we get (using the $x^1 = 0$)

$$\begin{aligned}\sum_{k\in[r]}\langle x^k - \bar{x}, v^k\rangle &= \frac{1}{2\eta}\left(-\|x^{r+1} - \bar{x}\|^2 + \|\bar{x}\|^2\right) + \frac{\eta}{2}\sum_{k\in[r]}\|v^k\|^2 \\ &\leqslant \frac{1}{2\eta}\|\bar{x}\|^2 + \frac{\eta}{2}\sum_{k\in[r]}\|v^k\|^2,\end{aligned}$$

which establishes the inequality (a).

To end the proof, it is enough to use the bounds $\|\bar{x}\| \leqslant B$ and $\|v^k\| \leqslant \rho$, and the value $B/\rho\sqrt{r}$ given to η: we get

$$\sum_{k\in[r]}\langle x^k - \bar{x}, v^k\rangle \leqslant B\rho\sqrt{r},$$

and the claim follows on dividing by r.

Remark. In next slide we use *Jensen's inequality*:

If $f : \mathcal{X} \to \mathbf{R}$ is convex, then

$$f(\lambda_1 x^1 + \cdots + \lambda_k x^k) \leqslant \lambda_1 f(x^1) + \cdots + \lambda_k f(x^k)$$

for any $x^1, \cdots, x^k \in \mathcal{X}$ and any $\lambda_1, \cdots, \lambda_k \in \mathbf{R}_+$ such that $\lambda_1 + \cdots + \lambda_k = 1$.

Proof. The statement is trivial for $k=1$, or if $\lambda_1=1$. So we may assume that $k\geqslant 2$ and $\lambda_1\neq 1$. Let

$$x'=(\lambda_2x^2+\cdots+\lambda_kx^k)/(1-\lambda_1).$$

Since $(\lambda_2+\cdots+\lambda_k)/(1-\lambda_1)=1$, $x'\in\mathcal{X}$ and hence

$$f(\lambda_1x^1+(1-\lambda_1)x')\leqslant\lambda_1f(x^1)+(1-\lambda_1)f(x').$$

By induction,

$$f(x')\leqslant\tfrac{\lambda_2}{1-\lambda_1}f(x^2)+\cdots+\tfrac{\lambda_k}{1-\lambda_k}f(x^k),$$

and the proof follows immeditely, as

$$(1-\lambda_1)f(x')\leqslant\lambda_2f(x^2)+\cdots+\lambda_kf(x^k).\qquad\square$$

Theorem. Let f be a convex ρ-Lipschitz function, and $\bar{x} = \mathrm{argmin}_{x:\|x\|\leqslant B} f(x)$. If we run the algorithm GD1 on f for r steps with $\eta = B/\rho\sqrt{r}$, then the output vector $\hat{x}$ satisfies

$$f(\hat{x}) - f(\bar{x}) \leqslant B\rho/\sqrt{r}.$$

Thus, for every $\epsilon > 0$, the inequality $f(\hat{x}) - f(\bar{x}) \leqslant \epsilon$ is achieved as soon as $r \geqslant B^2\rho^2/\epsilon^2$.

Proof. We have:

$$\begin{aligned}
f(\hat{x}) - f(\bar{x}) &= f\left(\tfrac{1}{r}\textstyle\sum_{k\in[r]} x^k\right) - f(\bar{x}) \quad \text{(defintion of } \hat{x}\text{)}\\
&\leqslant \tfrac{1}{r}\left(\textstyle\sum_{k\in[r]} f(x^k)\right) - f(\bar{x}) \quad \text{(Jensen's inequality)}\\
&= \tfrac{1}{r}\textstyle\sum_{k\in[r]} (f(x^k) - f(\bar{x}))\\
(*) \qquad &\leqslant \tfrac{1}{r}\textstyle\sum_{k\in[r]} \langle x^k - \bar{x}, \nabla f(x^k)\rangle \quad (f \text{ is convex})\\
&\leqslant B\rho/\sqrt{r}.
\end{aligned}$$

The last inequality is a consequence of $\|\nabla f(x^k)\| \leqslant \rho$ (Lemma on page 18) and the second part of the Lemma on page 24. □

The GD procedure works for nondifferentiable functions by using a subgradient of $f(x)$ at x^k.

The results on convergence remain the same.

The key point is that the inequality $(*)$ on the previous slide is valid for a subgradient s^k instead of $\nabla f(x^k)$.

1. **Input**: Initial value $x^0 = x^1$, η, μ
2. $x = x^1; p = x^1 - x^0$
3. **while not converged**:
4. $\quad x = x - \eta\nabla f(x) + \mu p$
5. $\quad p = \mu p - \eta\nabla f(x)$
6. $\quad$ **convergence check**
7. $\quad$ [**update** η], [**update** μ]
8. **return** x

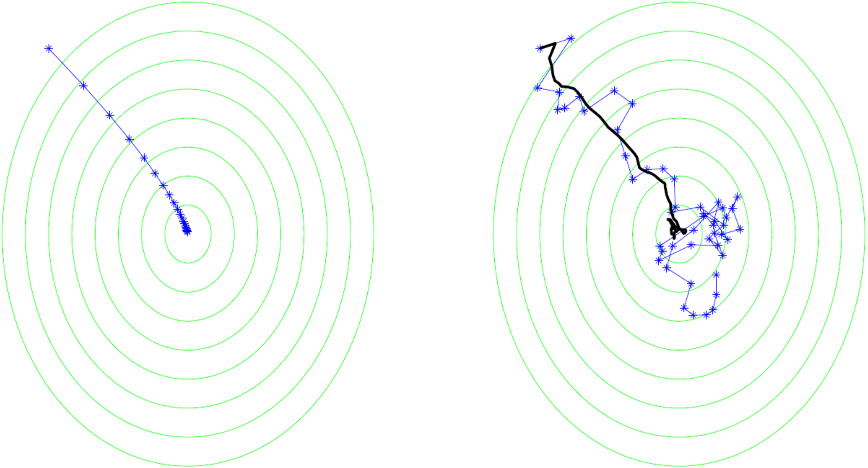
For comparisons of this GD3 (known as *heavy ball method* when η and μ are fixed) with GD1 and GD2, as well as with the *conjugate gradient method*, see [7, § 7.1]. See also § 7.2 for a short account of the Nestorov *accelerated gradient methods* and § 7.3 for *coordinate descent methods*.

Stochastic gradient descent

Stochastic gradients

Basic SGD algorithms

Convergence results



From Fig. 14.3 in [3], illustrating the behavior of the optimization steps when instead of the gradient an *stochastic gradient* is used, namely, a random vector whose *expected value* points in the same direction as the gradient.

Assume $\mathcal{H}$ is a hypothesis space of parameterized functions: $\{f_w\}_{w\in W}$. In algorithmic learning, the main problem is minimizing the loss (or risk) function $L(f_w) = L(w)$.

In empirical risk minimization, we used the empirical risk $L_{\mathcal{D}}(w)$, associated with data $\mathcal{D}$ to approximate $L(w)$. Notice that we cannot use gradient methods to directly minimize $L(w)$, as its definition depends on the unknown probability distribution ruling the generation of data.

The stochastic techniques allow to deal with the minimization of $L(w)$ by supplying a random vector v whose conditional expectated value is $\nabla L(w)$: $\mathbb{E}[v|w] = \nabla L(w)$.

For simplicity, assume first that the local loss function, $\ell(w, z)$ is differentiable. Then we can define the stochatic gradient, relative to w, as the random vector such that $\mathbb{E}[v|w] = \mathbb{E}_{z\sim\mathcal{P}}[\nabla_w \ell(w, z)]$. By linearity of the gradient,

$$\mathbb{E}_{z\sim\mathcal{P}}[\nabla_w \ell(w, z)] = \nabla_w \mathbb{E}_{z\sim\mathcal{P}}[\ell(w, z)] = \nabla L(w).$$

Thus $\nabla_w \ell(w, z)$ is an unbiased estimate of $\nabla L(w)$.

In practice this means sampling z and takinkg $\nabla_w \ell(w, z)$ as stochastic gradient at w.

For non-differentiable functions, $\nabla_w \ell(w, z)$ has to be replaced by a subgradient v of $\ell(w, z)$ at w. Then for any x we have $\ell(x, z) - \ell(w, z) \geqslant \langle x - w, v\rangle$ and taking expectation of both sides with respect to $z \sim \mathcal{P}$, we get

$$L(x) - L(w) \geqslant \mathbb{E}[\langle x - w, v\rangle] = \langle x - w, \mathbb{E}[v]\rangle,$$

which shows that $\mathbb{E}[v]$ is a subgradient of $L(w)$ at w.

1. **Parameters**: η (or $\eta_1, \eta_2, \ldots$) and r.
2. **require**: Initial value $w^1 = 0$
3. **for** $k = 1, 2, \cdots, r$
4. sample z
5. pick $v_k \in \partial\ell(w^k, z)$
6. update: $w^{k+1} = w^k - \eta v$
7. **return** $\bar{w} = \frac{1}{r}\sum_1^r w^k$

Appendix

Newton's method
Levenberg-Marquardt procedure

Let $\bar{x} = argmin_{x\in\mathcal{X}} f(x)$, $\mathcal{X}$ an open subset of $\mathbf{R}^n$. Assume that f is differentiable and let $\nabla^2 f(x) = Hf(x)$ be the *Hessian* of f, that is, the symmetric matrix $(\partial_i\partial_j f(x))_{i,j=1}^n$.

Newton's algorithm aims at approximating $\bar{x}$ starting with a *guess* x^0 and constructing a sequence $x^1, x^2, \cdots$ as follows:

$x^{k+1} = x^k + \Delta_k$, where $\Delta_k Hf(x^k) = -\nabla f(x^k)$.

The *heuristics* for this rule are:
(1) $\nabla f(x^{k+1}) \approx \nabla f(x^k) + (x^{k+1} - x^k)Hf(x^k)$;
(2) If $x^{k+1} = \bar{x}$, then we would have $0 = \nabla f(x^k) + (\bar{x} - x^k)Hf(x^k)$, wich would allow to find $\bar{x}$; and
(3) Proceed as if $\nabla f(x^{k+1}) = 0$ and replace $x^{k+1} - x^k$ by Δ_k, which leads to the equation $0 = \nabla f(x^k) + \Delta_k Hf(x^k)$.

Fact. $\|x^{k+1} - \bar{x}\| \leqslant C\|x^k - \bar{x}\|^2$.

This insures a fast convergence to $\bar{x}$ as soon as x^k is close to $\bar{x}$.

Levenberg-Marquardt for nonlinear least squares: combine gradient descent and Newton update rules into one rule, with a parameter λ. Small values of λ lean toward Newton, large values of λ will lean toward gradient descent.

One of the principal discoveries in machine learning in recent years is an empirical one—that *simple algorithms often suffice to solve difficult real-world learning problems*.

Machine learning algorithms generally arise via formulations as optimization problems, and, despite a massive classical toolbox of sophisticated optimization algorithms and a major modern effort to further develop that toolbox, the simplest algorithms— gradient descent, which dates to the 1840s [Cauchy, 1847] and stochastic gradient descent, which dates to the 1950s [Robbins and Monro, 1951]—*reign supreme in machine learning*.

References I

[1] H. Robbins and S. Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

[2] N. L. Roux, M. Schmidt, and F. R. Bach, "A stochastic gradient method with an exponential convergence-rate for finite training sets," *Advances in Neural Information Processing Systems*, pp. 2663–2671, 2012.

[3] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*.

Cambridge university press, 2014.

440 pp.

References II

[4] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd international conference for learning representations, San Diego*, ICLR, 2015.

https://arxiv.org/abs/1412.6980.

[5] S. Bubeck, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–358, 2015.

arXiv:1405.4980.

[6] R. H. Byrd, S. L. Hansen, J. Nocedal, and Y. Singer, "A stochastic quasi-Newton method for large-scale optimization," *SIAM Journal on Optimization*, vol. 26, no. 2, pp. 1008–1031, 2016.

https://arxiv.org/pdf/1401.7020.pdf.

References III

[7] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *Siam Review*, vol. 60, no. 2, pp. 223–311, 2018.

`https://arxiv.org/pdf/1606.04838.pdf`.

[8] G. Strang, *Linear algebra and learning from data*.

Wellesley-Cambridge Press, 2019.

[9] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, "On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points," 2019.

`https://arxiv.org/pdf/1902.04811.pdf`.

References IV

[10] C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan, "On nonconvex optimization for machine learning: Gradients, stochasticity, and saddle points," *Journal of the ACM (JACM)*, vol. 68, no. 2, pp. 1–29, 2021.

[11] J. Fearnley, P. W. Goldberg, A. Hollender, and R. Savani, "The complexity of gradient descent: CLS $=$ PPAD$\cap$PLS," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 46–59, 2021.