5.3 Stochastic gradient descent

- If we consider learning tasks with very large number of data $m \gg 1$, then the calculation of the gradients of empirical risk can become very costly:

$$\nabla \mathcal{R}_S(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{x}_i, y_i).$$

- **Example:** The HIGGS dataset for classification of Higgs boson-generating processes includes $m \approx 10^7$ training data with $d = 28$ features. Thus, in each step of the gradient procedure, we compute the mean of 10^7 gradient vectors in $\mathbb{R}^{28}$ ...

- Furthermore, in large data sets there are often many redundancies – not every data point contains new or unique information...

- This motivates the idea that one should not consider all training data per gradient step.

The stochastic gradient method uses exactly one, randomly selected, gradient $\nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{x}_i, y_i)$ per step:

Stochastic Gradient Descent (SGD)

Given a starting vector $\mathbf{w}_0 \in \mathbb{R}^p$ and an objective function

$$F(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} f_i(\mathbf{w}), \qquad f_i \colon \mathbb{R}^p \to \mathbb{R},$$

calculate for $k = 0, 1, 2, \ldots$ the iterates $\mathbf{w}_{k+1}$ as follows:

1. draw realization $i_k \in [m]$ of the uniformly random index variable

   uniform distribution

   $$I_k \sim \mathrm{U}([m]), \qquad [m] := \{1, \ldots, m\}$$

   where the I_k, $k \in \mathbb{N}$, are stochastically independent,

2. for a given deterministic step size $\eta_k > 0$, calculate

   $$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \underbrace{\nabla f_{i_k}(\mathbf{w}_k)}_{\text{only at that randomed uniquely choosed data}}.$$

Notes

- The generated sequence of iterates is random and, in particular, forms a Markov chain $(\mathbf{W}_k)_{k\in\mathbb{N}}$.
- The direction $-\nabla f_{i_k}(\mathbf{w}_k)$ is in general no longer a descent direction of F, but on average the SGD goes in the direction of the gradient:

$$\mathbb{E}[\mathbf{W}_{k+1} - \mathbf{W}_k \mid W_k] = \mathbb{E}_{\underbrace{\mathrm{U}([m])}_{\text{uniformly distributed}}}[-\underbrace{\eta_k}_{\text{step size}} \nabla f_{\underbrace{I_k}_{\text{Ranomly data}}}(\mathbf{W}_k)] = \boxed{-\eta_k \nabla F(\mathbf{W}_k)}$$

- The overhead of the method is independent of m, making it suitable for learning with very large datasets.
- Also regularized empirical risks $F(\mathbf{w}) = \lambda\|\mathbf{w}\|^2 + \mathcal{R}_S(\mathbf{w})$ follow the form $F(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{w})$:

$$\lambda\|\mathbf{w}\|^2 + \mathcal{R}_S(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^m \left(\lambda\|\mathbf{w}\|^2 \underline{\underline{+}}\, \ell(\mathbf{w}; \mathbf{x}_i, y_i)\right)$$

Example

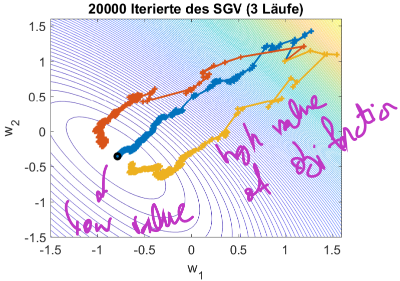
- We consider logistic regression for a dataset with $m = 19020$ data pairs and two parameters $\mathbf{w} = (w_1, w_2)$.

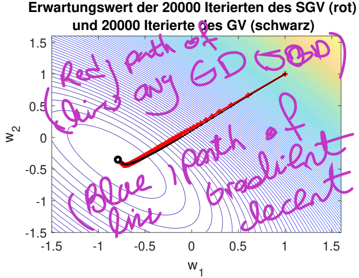
- I.e., for one step of gradient descent we can also do about 20000 steps of the SGD at about the same cost.

- We choose as initial value $\mathbf{w}_0 = (1, 1)$ and as step size $\eta_k = k^{-0.75}$.

- We consider 3 independent runs of the SGD (top) as well as the mean over 1000 runs (bottom).

- What do we observe?


20000 Iterierte des SGV (3 Läufe)


Erwartungswert der 20000 Iterierten des SGV (rot) und 20000 Iterierte des GV (schwarz)

- The SGD comes with approximately the same costs or the same number of evaluations of $\nabla_{\mathbf{w}} \ell(\mathbf{w}; \mathbf{x}_i, y_i)$ much closer to the actual minimum of F!

- In particular, we observe not only an apparent pathwise convergence of the SGD – that is, for each of the independent run – but ...

- ... also obtained by the deterministic sequence of the means of the iterates $\mathbb{E}[\mathbf{W}_k]$ a good approximation to the gradient descent.

- This is suggested a convergence analysis of the SGD in expectation:

$$\mathbb{E}[\mathbf{W}_k] \quad \text{respectively} \quad \mathbb{E}[F(\mathbf{W}_k)]\,.$$

- But before, we take another look at the SGD method...

Excursus: stochastic optimization

- The SGD can also be used to minimize objective function of the following form:

  $$F(\mathbf{w}) = \mathbb{E}_\nu[f(\mathbf{w};\xi)]\,,$$

  with differentiable $f\colon \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}$ and random variable $\xi \sim \nu$ in $\mathbb{R}^d$.

- The SGD then only requires that we can draw (independent) realizations ξ_k of ξ, and computes

  $$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta_k \nabla_{\mathbf{w}} f(\mathbf{w}_k;\xi_k)$$

- A calculation of the expected value w. r. t. ξ would yield the true gradient

  $$\nabla F(\mathbf{w}) = \mathbb{E}_\nu[\nabla_{\mathbf{w}} f(\mathbf{w};\xi)]$$

  but is in applications often not possible!

- Two special cases of stochastic optimization are now

  $$\mathcal{R}_\mu(\mathbf{w}) = \mathbb{E}_\mu[\ell(\mathbf{w}; \mathbf{X}, Y)],$$

  thus $f(\mathbf{w}; \xi) = \ell(\mathbf{w}; \mathbf{X}, Y)$ with $\xi = (\mathbf{X}, Y) \sim \mu$, and

  $$\mathcal{R}_s(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} \ell(\mathbf{w}, \mathbf{x}_i, y_i) = \mathbb{E}_{\mathrm{U}([m])}[\ell(\mathbf{w}; \mathbf{x}_I, y_I)].$$

  where $f(\mathbf{w}; \xi) = \ell(\mathbf{w}; \mathbf{x}_\xi, y_\xi)$ and $\xi \sim \mathrm{U}([m])$.

- I.e. also ERM tasks can be regarded as stochastic optimization with respect to the uniform distribution on the data...

- ... and theoretically even the minimization of the expected risk $\mathcal{R}_\mu$ per SGD is possible – if we can generate random data according to the (often unknown) distribution μ.

- The SGD dates back to H. Robbins and S. Monro, who developed and analyzed the method in an 8-page paper (1951).

Convergence analysis for SGD

GD is deterministic

SGD is Random,

- We focus again on strongly convex and L-smooth objective functions F.

it is exponentially faster than SGD

- In contrast to gradient descent, we do not have a guaranteed descent of the objective function (even on average):

Proposition 5.6:

Let $F\colon \mathbb{R}^p \to \mathbb{R}$ be given by $F = \frac{1}{m}\sum_{i=1}^m f_i$ and be L-smooth. Then, we have for the random iterates $\mathbf{W}_k$ of the SGD

were deterministic

when 0, what happens

$$\mathbb{E}[F(\mathbf{W}_{k+1}) - F(\mathbf{W}_k) \mid W_k] \leq -\eta_k \|\nabla F(\mathbf{W}_k)\|^2 + \frac{\eta_k^2\, L}{2}\mathbb{E}_{\mathrm{U}([m])}\left[\|\nabla f_{I_k}(\mathbf{W}_k)\|^2\right].$$

the certified negative in GD

sth like 0

if this is 0,

- The term $\mathbb{E}_{\mathrm{U}([m])}\left[\|\nabla f_{I_k}(\mathbf{w}_k)\|^2\right]$ is due to the variance of the estimator $\nabla f_{I_k}(\mathbf{w}_k)$ for the true gradient $\nabla F(\mathbf{w}_k)$.

$$* \quad \varphi(\mathbb{E}[Z]) \leq \mathbb{E}[\varphi(Z)]$$

proof Pro 5,6:

Since F is L-smooth we have $F(w_{k+1}) \le F(w_k) + \nabla F(w_k)^T (w_{k+1}-w_k) + \frac{L}{2}\|w_{k+1}-w_k\|^2$ $\boxed{*}$. Thus, we get: $\mathbb{E}[F(W_{k+1}) - F(W_k) \mid W_k]$

$$= \mathbb{E}_{I_k \sim u(n)}\left[F(W_k - \eta_k \nabla f_{I_k}(w_k)) - F(W_k)\right] \quad \boxed{**}$$

⇒ Hence using $\boxed{*}$, we have $\boxed{**}$. then,

$$\boxed{**} \le -\eta_k \mathbb{E}_{I_k \sim u[n]}\left[\nabla F(W_k)^T \nabla f_{I_k}(W_k)\right] + \frac{L}{2}\mathbb{E}_{I_k \sim u[n]}\left[\|\eta_k \nabla f_{I_k}(W_k)\|^2\right] = -\eta_k \underbrace{\|\nabla F(W_k)\|^2}_{\text{direction of Gradient}} + \frac{\eta_k^2 L}{2}\mathbb{E}_{I_k}\left[\|\nabla f_{I_k}(W_k)\|^2\right]$$

Theorem 5.7:

Let $F = \frac{1}{m}\sum_{i=1}^m f_i$ be λ-strongly convex and L-smooth. Further let for constants $a, b \in \mathbb{R}$ hold that

$$\frac{1}{m}\sum_{i=1}^m \|\nabla f_i(\mathbf{w})\|^2 \leq a + b\|\nabla F(\mathbf{w})\|^2 \qquad \forall \mathbf{w} \in \mathbb{R}^p.$$

(contral Red one via this in prev slide)

If the stepsizes η_k satisfy

$$\sum_{k=0}^{\infty} \eta_k = +\infty, \qquad \sum_{k=0}^{\infty} \eta_k^2 < +\infty,$$

should be finite

then we have for the iterates $\mathbf{W}_k$ of the SGD and $F^* := \min_{\mathbf{w}\in\mathbb{R}^p} F(\mathbf{w})$ that

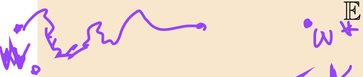
The avg of objective function would converge to Zero.

$$\lim_{k\to\infty} \mathbb{E}[F(\mathbf{W}_k)] = F^*.$$

(just convergence not fast)

In particular, for stepsizes $\eta_k = \frac{\beta}{\gamma+k}$ with $\beta \geq \frac{1}{\lambda}$, $\gamma \geq L\beta b$ we get

if $\sum \eta_k < \infty$



$$\mathbb{E}[F(\mathbf{W}_k) - \underline{F^*}] \leq \frac{\nu}{\gamma + k}, \qquad \text{where } \nu \geq \frac{\beta^2\, L\, a}{2(\beta\lambda - 1)}$$

) how fast convergen is.

$$\|\eta_k \nabla f_{I_k}(W_k)\| \leq K \eta_k$$

Proof Th 5,7:

By prop 5,6, and the assumptions we have $\mathbb{E}[F(w_{k+1}) - F(w_k) | w_k)$

$$\leq -\eta_k \|\nabla F(W_k)\|^2 + \frac{\eta_k^2 L}{2}\left(a + b\|\nabla + F(W_k)\|^2\right)$$

$$= \underbrace{\left(1 - \frac{Lb}{2}\eta_k\right)}_{\leq -\frac{1}{2}} \eta_k \|\nabla F(W_k)\|^2 + a\frac{L}{2}\eta_k^2$$

$\leq -\frac{1}{2}$ for sufficiently large k, since $\eta_k \to 0$.

$\leq -\frac{1}{2}\eta_k \|\nabla F(W_k)\|^2 + \frac{aL}{2}\eta_k^2$. using PLC we obtain after rearrangement $\mathbb{E}[F(w_{k+1}) - F^*] \leq (1 - \lambda\eta_k)\mathbb{E}[F(W_k) - F^*]$

$$+ \frac{aL}{2}\eta_k^2$$

choosing $\eta_k = \frac{\beta}{\gamma + k}$ we can prove via induction that:

$\mathbb{E}[F(W_k) - F^*] \leq \frac{\nu}{\gamma + k}$, since $\mathbb{E}[F(W_{k+1}) - F^*] \leq \left(1 - \lambda\frac{\beta}{\gamma + k}\right)\frac{\nu}{\gamma + k} + \frac{aL\beta^2}{2(\gamma + k)^2}$

$$= \frac{\gamma+k-1}{(\gamma+k)^2}\cdot V - \underbrace{\frac{\beta\lambda-1}{(\gamma+k)^2}V + \frac{\alpha\beta L}{2(\gamma+k)^2}}_{\leq 0 \text{ by assumption on } V.}$$

Hence we have $\leq \dfrac{\gamma+k-1}{(\gamma+k+1)(\gamma+k-1)}V \leq \dfrac{V}{\gamma+(k+1)}.$ #

Work complexity

- Even if the SGD (on expectation) shows slower convergence than gradient descent, it has the better work complexity.

- **Work complexity:** What computational effort is required to find the minimum of F up to an $\epsilon > 0$, i.e.,

$$\mathbb{E}[F(\mathbf{w}_k) - F(\mathbf{w}^*)] = F(\mathbf{w}_k) - F(\mathbf{w}^*) \leq \epsilon \ ?$$

- The gradient method requires only $\mathcal{O}(-\ln \epsilon)$ steps due to its exponential convergence.

- But per step we have an overhead of $\mathcal{O}(m)$ for computing the gradient – so a complexity of $\mathcal{O}(m \ln(1/\epsilon))$.

- For $\mathbb{E}[F(\mathbf{W}_k) - F(\mathbf{w}^*)] \leq \epsilon$ the SGD does require $\mathcal{O}(1/\epsilon)$ steps, but each with $\mathcal{O}(1)$ effort – thus a $\mathcal{O}(1/\epsilon)$ complexity.

SGD

- Hence, for large datasets $m \gg 1$ and moderate error bounds ϵ the SGD is preferable.

Extensions of SGD

- There are numerous extensions of the simple SGD. Many of these methods aim at a variance reduction for gradient estimation.

- For example, instead of taking one i_k, one could take a small batch of indices for gradient estimation – this reduces variance, but also increases effort – no improvement in complexity.

- More promising are methods that cleverly reuse already calculated gradients $\nabla f_{i_j}(\mathbf{w}_j)$ and thus aggregate old gradients to reduce the variance, e.g. SAGA.

  to reduce $\frac{\alpha L}{2} \eta_k^2$

- Thus, linear convergence can be achieved again, see Section 5.3 in Optimization Methods for Large-Scale Machine Learning.

- Overall, stochastic gradient methods are still a very recent research topic (ADAM is from 2014!).