Stochastic optimization. SGD. Variance reduction

Optimization in ML

Aleksandr Beznosikov

Skoltech

1 December 2023

Stochastic optimization: setting

- Consider the problem:

$$\min_{x \in \mathbb{R}^d} f(x).$$

Stochastic optimization: setting

- Consider the problem:

$$\min_{x \in \mathbb{R}^d} f(x).$$

- Let us now formulate the problem as follows:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \mathbb{E}_{\xi \sim \mathcal{D}}[f(x, \xi)] \right].$$

Stochastic optimization: setting

- Consider the problem:

$$\min_{x\in\mathbb{R}^d} f(x).$$

- Let us now formulate the problem as follows:

$$\min_{x\in\mathbb{R}^d} \left[f(x) := \mathbb{E}_{\xi\sim\mathcal{D}}[f(x,\xi)]\right].$$

- To understand the point, let's look at an example from machine learning:

$$\min_{x\in\mathbb{R}^d} \left[f(x) := \mathbb{E}_{\xi\sim\mathcal{D}}[\ell(g(x,\xi_x),\xi_y)]\right],$$

where $\mathcal{D}$ – data distribution (nature of the data), $\xi = (\xi_x, \xi_y)$ – sample: ξ_x – object (picture, text) and ξ_y – label, g – machine learning model (linear model, neural network), takes as input the object and customizable weights x, ℓ – loss function (penalizes the model for mismatches with the real label ξ_y).

Stochastic optimization: setting

- We want to «adjust» to nature, and to make the model losses on average over the whole distribution the smallest, i.e. the model best approximates the dependence of ξ_y on ξ_x.

Stochastic optimization: setting

- We want to «adjust» to nature, and to make the model losses on average over the whole distribution the smallest, i.e. the model best approximates the dependence of ξ_y on ξ_x.
- **Question:** what's the problem?

Stochastic optimization: setting

- We want to «adjust» to nature, and to make the model losses on average over the whole distribution the smallest, i.e. the model best approximates the dependence of ξ_y on ξ_x.
- **Question:** what's the problem? the function f (as well as gradients and higher derivatives) don't co because we don't know $\mathcal{D}$ (that's the essence of approximating something complex and unknown), and even if we do, the integral (expectation) is often not so easy to take.

Stochastic optimization: setting

- We want to «adjust» to nature, and to make the model losses on average over the whole distribution the smallest, i.e. the model best approximates the dependence of ξ_y on ξ_x.
- **Question:** what's the problem? the function f (as well as gradients and higher derivatives) don't co because we don't know $\mathcal{D}$ (that's the essence of approximating something complex and unknown), and even if we do, the integral (expectation) is often not so easy to take.
- We need a method that can handle $\nabla f(x, \xi)$ (gradient of a particular sample from the data distribution). That is, we want to work in online mode: samples come in, we process them (we can read the gradient).

Stochastic optimization: setting

- We want to «adjust» to nature, and to make the model losses on average over the whole distribution the smallest, i.e. the model best approximates the dependence of ξ_y on ξ_x.
- **Question:** what's the problem? the function f (as well as gradients and higher derivatives) don't co because we don't know $\mathcal{D}$ (that's the essence of approximating something complex and unknown), and even if we do, the integral (expectation) is often not so easy to take.
- We need a method that can handle $\nabla f(x, \xi)$ (gradient of a particular sample from the data distribution). That is, we want to work in online mode: samples come in, we process them (we can read the gradient).
- The natural assumption is that the data is unbiased:

$$\mathbb{E}_{\xi \sim \mathcal{D}}[\nabla f(x, \xi)] = \nabla f(x).$$

Stochastic optimization: a different setting

- Often in machine learning we do not start from «zero» and a training sample is given, then often the learning problem is written in the form of minimizing empirical risk:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{n} \sum_{i=1}^{n} [\ell(g(x, \xi_{x,i}), \xi_{y,i})] \right],$$

where $\{\xi_i\}_{i=1}^n$ is a sample from $\mathcal{D}$, g is a model, ℓ is a function. This formulation is called offline (the data is fixed, not real-time).

Stochastic optimization: a different setting

- Often in machine learning we do not start from «zero» and a training sample is given, then often the learning problem is written in the form of minimizing empirical risk:

$$\min_{x\in\mathbb{R}^d}\left[f(x) := \frac{1}{n}\sum_{i=1}^{n}[\ell(g(x,\xi_{x,i}),\xi_{y,i})]\right],$$

where $\{\xi_i\}_{i=1}^n$ is a sample from $\mathcal{D}$, g is a model, ℓ is a function. This formulation is called offline (the data is fixed, not real-time).
- **Question:** what's the relationship between emphonline and offline?

Stochastic optimization: a different setting

- Often in machine learning we do not start from «zero» and a training sample is given, then often the learning problem is written in the form of minimizing empirical risk:

$$\min_{x \in \mathbb{R}^d} \left[f(x) := \frac{1}{n} \sum_{i=1}^{n} [\ell(g(x, \xi_{x,i}), \xi_{y,i})] \right],$$

where $\{\xi_i\}_{i=1}^n$ is a sample from $\mathcal{D}$, g is a model, ℓ is a function. This formulation is called offline (the data is fixed, not real-time).

- **Question:** what's the relationship between emphonline and offline? offline is a Monte Carlo approximation of the original integral (expectation matrix). If the number of samples is big, the approximation via finite sum will tend to the real integral (under certain assumptions).

Stochastic optimization: a different setting

- **Question:** is it already possible to read the gradient in the offline setting?

Stochastic optimization: a different setting

- **Question:** is it already possible to read the gradient in the offline setting? yes! It turns out that the problem is solved and the lecture is over. But for some reason machine learning often doesn't use full honest gradients. **Question:** why?

Stochastic optimization: a different setting

- **Question:** is it already possible to read the gradient in the offline setting? yes! It turns out that the problem is solved and the lecture is over. But for some reason machine learning often doesn't use full honest gradients. **Question:** why? it's expensive/expensive to do a full gradient.

Stochastic optimization: a different setting

- **Question:** is it already possible to read the gradient in the offline setting? yes! It turns out that the problem is solved and the lecture is over. But for some reason machine learning often doesn't use full honest gradients. **Question:** why? it's expensive/expensive to do a full gradient.
- So instead of a full gradient, a random sample gradient is called:

$$\nabla f(x, \xi_i), \quad \text{где} \quad \xi_i \text{ generated independently and uniformly from } \mathcal{D} \text{ or } [n].$$

Stochastic gradient descent

- Simple idea – modify the gradient descent again and see what happens.

Algorithm 1 SGD

Input: stepsize $\{\gamma_k\}_{k=0} > 0$, starting point $x^0 \in \mathbb{R}^d$, number of iterations K

1: **for** $k = 0, 1, \dots, K-1$ **do**

2: Generate independently ξ^k

3: Compute stochastic gradient $\nabla f(x^k, \xi^k)$

4: $x^{k+1} = x^k - \gamma_k \nabla f(x^k, \xi^k)$

5: **end for**

Output: x^K

Conditional mathematical expectation

- The convergence proof will require the introduction of a conditional mathematical expectation:

$$\mathbb{E}\left[\cdot \mid x^k\right] = \mathbb{E}\left[\cdot \mid \mathcal{F}_k\right],$$

where $\mathcal{F}_k$ – σ-algebra generated by x^0, $\xi^0, \ldots, \xi^{k-1}$.

Conditional mathematical expectation

- The convergence proof will require the introduction of a conditional mathematical expectation:

$$\mathbb{E}\left[\cdot \mid x^k\right] = \mathbb{E}\left[\cdot \mid \mathcal{F}_k\right],$$

where $\mathcal{F}_k$ – σ-algebra generated by x^0, $\xi^0, \ldots, \xi^{k-1}$.

- The point – «to fix» all randomness that occurred before k iteration and expect only on randomness that remains unfrozen.

Conditional mathematical expectation

- The convergence proof will require the introduction of a conditional mathematical expectation:

$$\mathbb{E}\left[\cdot \mid x^k\right] = \mathbb{E}\left[\cdot \mid \mathcal{F}_k\right],$$

where $\mathcal{F}_k$ – σ-algebra generated by x^0, $\xi^0, \ldots, \xi^{k-1}$.
- The point – «to fix» all randomness that occurred before k iteration and expect only on randomness that remains unfrozen. **Question:** such a mathematical expectation gives an output: something deterministic or random?

Conditional mathematical expectation

- The convergence proof will require the introduction of a conditional mathematical expectation:

$$\mathbb{E}\left[\cdot \mid x^k\right] = \mathbb{E}\left[\cdot \mid \mathcal{F}_k\right],$$

where $\mathcal{F}_k$ – σ-algebra generated by x^0, $\xi^0, \ldots, \xi^{k-1}$.

- The point – «to fix» all randomness that occurred before k iteration and expect only on randomness that remains unfrozen. **Question:** such a mathematical expectation gives an output: something deterministic or random? Random, depending on random variables x^0, $\xi^0, \ldots, \xi^{k-1}$.

Conditional mathematical expectation

- The convergence proof will require the introduction of a conditional mathematical expectation:

$$\mathbb{E}\left[\cdot \mid x^k\right] = \mathbb{E}\left[\cdot \mid \mathcal{F}_k\right],$$

where $\mathcal{F}_k$ – σ-algebra generated by x^0, $\xi^0, \ldots, \xi^{k-1}$.
- The point – «to fix» all randomness that occurred before k iteration and expect only on randomness that remains unfrozen. **Question:** such a mathematical expectation gives an output: something deterministic or random? Random, depending on random variables x^0, $\xi^0, \ldots, \xi^{k-1}$.
- We will also need the law of total mathematical expectation (tower property):

$$\mathbb{E}\left[\mathbb{E}[X|Y]\right] = \mathbb{E}[X].$$

Convergence: a proof

- We will prove in the case when f is L-smooth and μ-simply convex.
- We also introduce new assumptions concerning the stochastic gradient:

$$\mathbb{E}_\xi[\nabla f(x,\xi)] = \nabla f(x), \quad \mathbb{E}_\xi\left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] \leq \sigma^2.$$

- Let us start as before:

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\gamma_k\langle \nabla f(x^k,\xi^k), x^k - x^*\rangle + \gamma_k^2\|\nabla f(x^k,\xi^k)\|^2.$$

Convergence: a proof

- We will prove in the case when f is L-smooth and μ-simply convex.
- We also introduce new assumptions concerning the stochastic gradient:

$$\mathbb{E}_\xi[\nabla f(x,\xi)] = \nabla f(x), \quad \mathbb{E}_\xi\left[\|\nabla f(x,\xi) - \nabla f(x)\|_2^2\right] \leq \sigma^2.$$

- Let us start as before:

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\gamma_k\langle\nabla f(x^k,\xi^k), x^k - x^*\rangle + \gamma_k^2\|\nabla f(x^k,\xi^k)\|^2.$$

- We take the conditional mat expectation by randomness only at iteration k (it is important that x^k – is a non-random variable with respect to the conditional m.o.):

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] = \|x^k - x^*\|^2 - 2\gamma_k\langle\mathbb{E}\left[\nabla f(x^k,\xi^k) \mid x^k\right], x^k - x^*\rangle + \gamma_k^2\mathbb{E}\left[\|\nabla f(x^k,\xi^k)\|^2 \mid x^k\right].$$

Convergence: a proof

- Work with $\mathbb{E}\left[\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k\right]$:

$$
\begin{aligned}
\mathbb{E}\left[\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k\right] &= \langle \mathbb{E}\left[\nabla f(x^k, \xi^k) \mid x^k\right], x^k - x^* \rangle \\
&= \langle \nabla f(x^k, \xi^k), x^k - x^* \rangle
\end{aligned}
$$

Convergence: a proof

- Work with $\mathbb{E}\left[\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k\right]$:

$$\begin{aligned}\mathbb{E}\left[\langle \nabla f(x^k, \xi^k), x^k - x^* \rangle \mid x^k\right] &= \left\langle \mathbb{E}\left[\nabla f(x^k, \xi^k) \mid x^k\right], x^k - x^* \right\rangle \\ &= \langle \nabla f(x^k, \xi^k), x^k - x^* \rangle\end{aligned}$$

- Work with $\mathbb{E}\left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k\right]$:

$$\begin{aligned}\mathbb{E}\left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k\right] =& \mathbb{E}\left[\left\|\nabla f(x^k, \xi^k) - \nabla f(x^k) + \nabla f(x^k)\right\|^2 \mid x^k\right] \\ =& \mathbb{E}\left[\left\|\nabla f(x^k, \xi^k) - \nabla f(x^k)\right\|^2 \mid x^k\right] \\ &+ \mathbb{E}\left[\left\|\nabla f(x^k)\right\|^2 \mid x^k\right] \\ &+ 2\mathbb{E}\left[\langle \nabla f(x^k, \xi^k) - \nabla f(x^k), \nabla f(x^k) \rangle \mid x^k\right].\end{aligned}$$

Convergence: a proof

- Continuing:

$$\mathbb{E}\left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k\right] = \mathbb{E}\left[\left\|\nabla f(x^k, \xi^k) - \nabla f(x^k)\right\|^2 \mid x^k\right] + \left\|\nabla f(x^k)\right\|^2 + 2\langle \mathbb{E}\left[\nabla f(x^k, \xi^k) \mid x^k\right] - \nabla f(x^k), \nabla f(x^k)\rangle.$$

Convergence: a proof

- Continuing:

$$\mathbb{E}\left[\|\nabla f(x^k,\xi^k)\|^2 \mid x^k\right] = \mathbb{E}\left[\left\|\nabla f(x^k,\xi^k) - \nabla f(x^k)\right\|^2 \mid x^k\right] + \left\|\nabla f(x^k)\right\|^2 + 2\langle \mathbb{E}\left[\nabla f(x^k,\xi^k) \mid x^k\right] - \nabla f(x^k), \nabla f(x^k)\rangle.$$

- The stochastic gradient assumption gives

$$\mathbb{E}\left[\|\nabla f(x^k,\xi^k)\|^2 \mid x^k\right] \leq \sigma^2 + \left\|\nabla f(x^k)\right\|^2.$$

Convergence: a proof

- Everything we got:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] = \|x^k - x^*\|^2 - 2\gamma_k \langle \mathbb{E}\left[\nabla f(x^k, \xi^k) \mid x^k\right], x^k - x^* \rangle + \gamma_k^2 \mathbb{E}\left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k\right].$$

$$\mathbb{E}\left[\nabla f(x^k, \xi^k) \mid x^k\right] = \nabla f(x^k).$$

$$\mathbb{E}\left[\|\nabla f(x^k, \xi^k)\|^2 \mid x^k\right] \leq \sigma^2 + \left\|\nabla f(x^k)\right\|^2.$$

Convergence: a proof

- Everything we got:

$$\mathbb{E}\left[\|x^{k+1}-x^*\|^2 \mid x^k\right] =\|x^k-x^*\|^2 - 2\gamma_k\langle \mathbb{E}\left[\nabla f(x^k,\xi^k) \mid x^k\right], x^k - x^*\rangle + \gamma_k^2 \mathbb{E}\left[\|\nabla f(x^k,\xi^k)\|^2 \mid x^k\right].$$

$$\mathbb{E}\left[\nabla f(x^k,\xi^k) \mid x^k\right] = \nabla f(x^k).$$

$$\mathbb{E}\left[\|\nabla f(x^k,\xi^k)\|^2 \mid x^k\right] \leq \sigma^2 + \left\|\nabla f(x^k)\right\|^2.$$

- Finally,

$$\mathbb{E}\left[\|x^{k+1}-x^*\|^2 \mid x^k\right] \leq\|x^k-x^*\|^2 - 2\gamma_k\langle \nabla f(x^k), x^k - x^*\rangle + \gamma_k^2\|\nabla f(x^k)\|^2 + \gamma_k^2\sigma^2.$$

Convergence: a proof

- Then there's the usual: L-smoothness and μ-strong convexity.

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] \leq & \|x^k - x^*\|^2 - 2\gamma_k \left(f(x^k) - f(x^*) + \frac{\mu}{2}\|x^k - x^*\|_2^2\right) \\
& + 2\gamma_k^2 L(f(x^k) - f(x^*)) + \gamma_k^2\sigma^2 \\
= & (1 - \gamma_k\mu)\|x^k - x^*\|^2 + \gamma_k^2\sigma^2 \\
& - 2\gamma_k(1 - \gamma_k L)(f(x^k) - f(x^*)).
\end{aligned}
$$

- If $\gamma_k \leq \frac{1}{L}$, then

$$
\mathbb{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] \leq (1 - \gamma_k\mu)\|x^k - x^*\|^2 + \gamma_k^2\sigma^2.
$$

Convergence: a proof

- Then there's the usual: L-smoothness and μ-strong convexity.

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1}-x^*\|^2 \mid x^k\right] \leq & \|x^k-x^*\|^2 - 2\gamma_k\left(f(x^k)-f(x^*)+\frac{\mu}{2}\|x^k-x^*\|_2^2\right) \\
& + 2\gamma_k^2 L(f(x^k)-f(x^*)) + \gamma_k^2\sigma^2 \\
= & (1-\gamma_k\mu)\|x^k-x^*\|^2 + \gamma_k^2\sigma^2 \\
& - 2\gamma_k(1-\gamma_k L)(f(x^k)-f(x^*)).
\end{aligned}
$$

- If $\gamma_k \leq \frac{1}{L}$, then

$$
\mathbb{E}\left[\|x^{k+1}-x^*\|^2 \mid x^k\right] \leq (1-\gamma_k\mu)\|x^k-x^*\|^2 + \gamma_k^2\sigma^2.
$$

- Taking the full expectation and applying tower property:

$$
\mathbb{E}\left[\|x^{k+1}-x^*\|^2\right] \leq (1-\gamma_k\mu)\mathbb{E}\left[\|x^k-x^*\|^2\right] + \gamma_k^2\sigma^2.
$$

SGD convergence

Theorem

Let the unconditional stochastic optimization problem with L-smooth, μ-strongly convex objective function f be solved using SGD with $\gamma_k \leq \frac{1}{L}$ under saturation and boundedness of the variance of the stochastic gradient. Then the following convergence estimate is valid

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq (1 - \gamma_k \mu)\mathbb{E}\left[\|x^k - x^*\|^2\right] + \gamma_k^2 \sigma^2.$$

SGD convergence: an analysis

- Constant stepsize $\gamma_k \equiv \gamma$, then

$$
\begin{aligned}
\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq & (1-\gamma\mu)\mathbb{E}\left[\|x^{k-1} - x^*\|^2\right] + \gamma^2\sigma^2 \\
\leq & (1-\gamma\mu)^2\mathbb{E}\left[\|x^{k-2} - x^*\|^2\right] \\
& + (1-\gamma\mu)\gamma^2\sigma^2 + \gamma^2\sigma^2 \\
\leq & \ldots \\
\leq & (1-\gamma\mu)^k\mathbb{E}\left[\|x^0 - x^*\|^2\right] + \gamma^2\sigma^2\sum_{i=0}^{k-1}(1-\gamma\mu)^i.
\end{aligned}
$$

SGD convergence: an analysis

- Constant stepsize $\gamma_k \equiv \gamma$, then

$$
\begin{aligned}
\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq & (1-\gamma\mu)\mathbb{E}\left[\|x^{k-1} - x^*\|^2\right] + \gamma^2\sigma^2 \\
\leq & (1-\gamma\mu)^2\mathbb{E}\left[\|x^{k-2} - x^*\|^2\right] \\
& + (1-\gamma\mu)\gamma^2\sigma^2 + \gamma^2\sigma^2 \\
\leq & \ldots \\
\leq & (1-\gamma\mu)^k\mathbb{E}\left[\|x^0 - x^*\|^2\right] + \gamma^2\sigma^2\sum_{i=0}^{k-1}(1-\gamma\mu)^i.
\end{aligned}
$$

- **Question:** how to evaluate the second summand?

SGD convergence: an analysis

- Constant stepsize $\gamma_k \equiv \gamma$, then

$$
\begin{aligned}
\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq & (1-\gamma\mu)\mathbb{E}\left[\|x^{k-1} - x^*\|^2\right] + \gamma^2\sigma^2 \\
\leq & (1-\gamma\mu)^2\mathbb{E}\left[\|x^{k-2} - x^*\|^2\right] \\
& + (1-\gamma\mu)\gamma^2\sigma^2 + \gamma^2\sigma^2 \\
\leq & \ldots \\
\leq & (1-\gamma\mu)^k\mathbb{E}\left[\|x^0 - x^*\|^2\right] + \gamma^2\sigma^2\sum_{i=0}^{k-1}(1-\gamma\mu)^i.
\end{aligned}
$$

- **Question:** how to evaluate the second summand? Geometric progression: $\sum_{i=0}^{k-1}(1-\gamma\mu)^i \leq \sum_{i=0}^{+\infty}(1-\gamma\mu)^i = \frac{1}{\gamma\mu}$:

$$
\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq (1-\gamma\mu)^k\mathbb{E}\left[\|x^0 - x^*\|^2\right] + \frac{\gamma\sigma^2}{\mu}.
$$

Convergence of SGD: an analysis

- The result:

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq (1-\gamma\mu)^k \mathbb{E}\left[\|x^0 - x^*\|^2\right] + \frac{\gamma\sigma^2}{\mu},$$

  is similar to what we've already seen for gradient descent.
- The first term – linear convergence to the solution

Convergence of SGD: an analysis

- The result:

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq (1-\gamma\mu)^k \mathbb{E}\left[\|x^0 - x^*\|^2\right] + \frac{\gamma\sigma^2}{\mu},$$

  is similar to what we've already seen for gradient descent.
- The first term – linear convergence to the solution
- The second term – indicates that some precision (depending on γ, σ and μ) the method cannot overcome and starts oscillating, no longer approaching the solution.

Convergence of SGD: problems

How can we try to solve problems of inexact convergence?

Convergence of SGD: problems

How can we try to solve problems of inexact convergence?

- Reduce the step. For example, take $\gamma_k = \frac{1}{k+1}$ или $\gamma_k = \frac{1}{\sqrt{k+1}}$.
  Question: what is the plus and minus view?

Convergence of SGD: problems

How can we try to solve problems of inexact convergence?

- Reduce the step. For example, take $\gamma_k = \frac{1}{k+1}$ или $\gamma_k = \frac{1}{\sqrt{k+1}}$.
  Question: what is the plus and minus view? Plus – more precisely convergence, minus – loss of linear convergence at the beginning.

Convergence of SGD: problems

How can we try to solve problems of inexact convergence?

- Reduce the step. For example, take $\gamma_k = \frac{1}{k+1}$ или $\gamma_k = \frac{1}{\sqrt{k+1}}$. **Question:** what is the plus and minus view? Plus – more precisely convergence, minus – loss of linear convergence at the beginning.
- Reduce σ. **Question:** how?

Convergence of SGD: problems

How can we try to solve problems of inexact convergence?

- Reduce the step. For example, take $\gamma_k = \frac{1}{k+1}$ или $\gamma_k = \frac{1}{\sqrt{k+1}}$.
  Question: what is the plus and minus view? Plus – more precisely convergence, minus – loss of linear convergence at the beginning.
- Reduce σ. **Question:** how? With the batching technique:

$$\nabla f(x^k, \xi^k) \quad \to \quad \frac{1}{b} \sum_{j \in S^k} \nabla f(x, \xi_j),$$

where S^k is the set of indices from $[n]$, $|S^k| = b$, and all indices are generated independently of each other.

Convergence of SGD: batching

- **Question:** what can we say about

$$\mathbb{E}\left[\frac{1}{b}\sum_{j \in S^k} \nabla f(x, \xi_j) \mid x^k\right], \quad \mathbb{E}\left[\left\|\frac{1}{b}\sum_{j \in S^k} (\nabla f(x, \xi_j) - \nabla f(x))\right\|_2^2 \mid x^k\right]?$$

Convergence of SGD: batching

- **Question:** what can we say about

$$\mathbb{E}\left[\frac{1}{b}\sum_{j\in S^k}\nabla f(x,\xi_j) \mid x^k\right], \quad \mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in S^k}(\nabla f(x,\xi_j) - \nabla f(x))\right\|_2^2 \mid x^k\right]?$$

- Independence gives

$$\mathbb{E}\left[\frac{1}{b}\sum_{j\in S^k}\nabla f(x,\xi_j) \mid x^k\right] = \nabla f(x),$$

$$\mathbb{E}\left[\left\|\frac{1}{b}\sum_{j\in S^k}(\nabla f(x,\xi_j) - \nabla f(x))\right\|_2^2 \mid x^k\right] \le \frac{\sigma^2}{b}$$

- It turns out that the variance can be reduced by a factor of b, but then the computation of the stochastic gradient becomes more expensive.

Convergence of SGD

- As a result, we can select a strategy for selecting steps and achieve the following convergence estimate:

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathbb{E}\left[\|x^0 - x^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

Linear on the «deterministic» part and sublinear on the «stochastic» part.

Convergence of SGD

- As a result, we can select a strategy for selecting steps and achieve the following convergence estimate:

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathbb{E}\left[\|x^0 - x^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

Linear on the «deterministic» part and sublinear on the «stochastic» part.

- Nesterov's acceleration is possible:

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathbb{E}\left[\|x^0 - x^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

Convergence of SGD

- As a result, we can select a strategy for selecting steps and achieve the following convergence estimate:

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathbb{E}\left[\|x^0 - x^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

Linear on the «deterministic» part and sublinear on the «stochastic» part.

- Nesterov's acceleration is possible:

$$\mathbb{E}\left[\|x^k - x^*\|^2\right] \leq \left(1 - \sqrt{\frac{\mu}{L}}\right)^k \mathbb{E}\left[\|x^0 - x^*\|^2\right] + \frac{\sigma^2}{\mu^2 bk}.$$

The important detail is that only the first term is improved/accelerated, the second term (which is due to stochasticity) remains the same. It turns out that it cannot be changed and the result above is optimal.

Why does SGD not converge?

- Initially SGD with constant has the same behavior as gradient descent: $x \to x^*$, but then oscillations start. **Question:** what is this related to? What is so unpleasant about the physics of the method?

Why does SGD not converge?

- Initially SGD with constant has the same behavior as gradient descent: $x \to x^*$, but then oscillations start. **Question:** what is this related to? What is so unpleasant about the physics of the method? In gradient descent, $\nabla f(x) \to \nabla f(x^*) = 0$. Now nobody guarantees this: $\nabla f(x, \xi)$ may not tend to 0.

Why does SGD not converge?

- Initially SGD with constant has the same behavior as gradient descent: $x \to x^*$, but then oscillations start. **Question:** what is this related to? What is so unpleasant about the physics of the method? In gradient descent, $\nabla f(x) \to \nabla f(x^*) = 0$. Now nobody guarantees this: $\nabla f(x, \xi)$ may not tend to 0.
- This is explainable in machine learning: x^* – minimizes the loss over the whole sample/over the whole distribution. $f(x, \xi)$ reflects only the loss on the sample ξ. No one guarantees that x^* – the best model setting for a particular sample ξ.

Why does SGD not converge?

- Initially SGD with constant has the same behavior as gradient descent: $x \to x^*$, but then oscillations start. **Question:** what is this related to? What is so unpleasant about the physics of the method? In gradient descent, $\nabla f(x) \to \nabla f(x^*) = 0$. Now nobody guarantees this: $\nabla f(x, \xi)$ may not tend to 0.
- This is explainable in machine learning: x^* – minimizes the loss over the whole sample/over the whole distribution. $f(x, \xi)$ reflects only the loss on the sample ξ. No one guarantees that x^* – the best model setting for a particular sample ξ.
- Because of the fact that in the general case $\nabla f(x^*, \xi) \neq 0$ for some ξ and the oscillatory effect occurs.

Modifying SGD

- The idea – to take a method like SGD:

$$x^{k+1} = x^k - \gamma g^k,$$

where

$$g^k \to \nabla f(x^*) = 0, \quad \text{if} \quad x^k \to x^*.$$

Whenever possible:

$$\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k) \quad \text{or} \quad \mathbb{E}\left[g^k\right] = \nabla f(x^k).$$

Modifying SGD

- The idea – to take a method like SGD:

$$x^{k+1} = x^k - \gamma g^k,$$

where

$$g^k \to \nabla f(x^*) = 0, \quad \text{if} \quad x^k \to x^*.$$

Whenever possible:

$$\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k) \quad \text{or} \quad \mathbb{E}\left[g^k\right] = \nabla f(x^k).$$

- In the general online case this is not realizable. But it is possible in the offline case:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x),$$

where, generating uniformly and independently i_k.

SAGA

Algorithm 2 SAGA

Input: step $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, memory $y_i^0 = 0$ for all $i \in [n]$, number of iterations K

1: **for** $k = 0, 1, \ldots, K-1$ **do**

2: Generate independetly i_k

3: Вычислить $g^k = \nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n}\sum\limits_{j=1}^{n} y_j^k$

4: Update $y_i^{k+1} = \begin{cases} \nabla f_i(x^k), & \text{если } i = i_k \\ y_i^k, & \text{elsewhere} \end{cases}$

5: $x^{k+1} = x^k - \gamma g^k$

6: **end for**

Output: x^K

SAGA

- Idea – if I once counted the gradient for f_i, why forget it? Let's keep it!

SAGA

- Idea – if I once counted the gradient for f_i, why forget it? Let's keep it!
- $\frac{1}{n}\sum\limits_{j=1}^{n} y_j^k$ – «delayed» version $\nabla f(x^k)$.

SAGA

- Idea – if I once counted the gradient for f_i, why forget it? Let's keep it!
- $\frac{1}{n}\sum\limits_{j=1}^{n} y_j^k$ – «delayed» version $\nabla f(x^k)$.
- $\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$.

SAGA

- Idea – if I once counted the gradient for f_i, why forget it? Let's keep it!
- $\frac{1}{n}\sum\limits_{j=1}^{n} y_j^k$ – «delayed» version $\nabla f(x^k)$.
- $\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$.
- If $x^k \to x^*$, we have $y_j^k \to \nabla f_j(x^*)$, and $\frac{1}{n}\sum\limits_{j=1}^{n} y_j^k \to \nabla f(x^*) = 0$.

  Therefore, $g^k \to 0$.

SAGA

- Idea – if I once counted the gradient for f_i, why forget it? Let's keep it!
- $\frac{1}{n}\sum\limits_{j=1}^{n} y_j^k$ – «delayed» version $\nabla f(x^k)$.
- $\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$.
- If $x^k \to x^*$, we have $y_j^k \to \nabla f_j(x^*)$, and $\frac{1}{n}\sum\limits_{j=1}^{n} y_j^k \to \nabla f(x^*) = 0$.

  Therefore, $g^k \to 0$.
- On the downside: extra $\mathcal{O}(nd)$ memory.

SAGA: a proof

- All f_i are L-smooth and convex, and $f - \mu$ is strongly convex.

SAGA: a proof

- All f_i are L-smooth and convex, and $f - \mu$ is strongly convex.
- Know these steps:

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\gamma\langle g^k, x^k - x^*\rangle + \gamma^2\|g^k - \nabla f(x^*)\|_2^2.$$

SAGA: a proof

- All f_i are L-smooth and convex, and $f - \mu$ is strongly convex.
- Know these steps:

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\gamma\langle g^k, x^k - x^*\rangle + \gamma^2\|g^k - \nabla f(x^*)\|_2^2.$$

- We take the conditional mat expectation at iteration k:

$$\begin{aligned}\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] =& \|x^k - x^*\|_2^2 - 2\gamma\langle \mathbb{E}\left[g^k \mid x^k\right], x^k - x^*\rangle \\ &+ \gamma^2\mathbb{E}\left[\|g^k - \nabla f(x^*)\|_2^2 \mid x^k\right].\end{aligned}$$

SAGA: a proof

- Work with $\mathbb{E}\left[g^k \mid x^k\right]$:

$$
\begin{aligned}
\mathbb{E}\left[g^k \mid x^k\right] &= \mathbb{E}\left[\nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n}\sum_{j=1}^{n} y_j^k \mid x^k\right] \\
&= \mathbb{E}\left[\nabla f_{i_k}(x^k) - y_{i_k}^k \mid x^k\right] + \frac{1}{n}\sum_{j=1}^{n} y_j^k \\
&= \frac{1}{n}\sum_{j=1}^{n}\left[\nabla f_j(x^k) - y_j^k\right] + \frac{1}{n}\sum_{j=1}^{n} y_j^k \\
&= \nabla f(x^k)
\end{aligned}
$$

SAGA: a proof

- Work with $\mathbb{E}\left[\|g^k - \nabla f(x^*)\|_2^2 \mid x^k\right]$:

$$
\begin{aligned}
\mathbb{E}\left[\|g^k - \nabla f(x^*)\|_2^2 \mid x^k\right] =& \mathbb{E}\left[\|\nabla f_{i_k}(x^k) - y_{i_k}^k + \frac{1}{n}\sum_{j=1}^n y_j^k - \nabla f(x^*)\|_2^2 \mid x^k\right] \\
=& \mathbb{E}\Bigg[\|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*) + \nabla f_{i_k}(x^*) - y_{i_k}^k \\
& + \frac{1}{n}\sum_{j=1}^n y_j^k - \nabla f(x^*)\|_2^2 \mid x^k\Bigg] \\
\leq& 2\mathbb{E}\left[\|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)\|_2^2 \mid x^k\right] \\
& + 2\mathbb{E}\left[\|\nabla f_{i_k}(x^*) - y_{i_k}^k + \frac{1}{n}\sum_{j=1}^n y_j^k - \nabla f(x^*)\|_2^2 \mid x^k\right]
\end{aligned}
$$

SAGA: a proof

- Using that $\mathbb{D}\xi \leq \mathbb{E}[\xi^2]$:

$$
\begin{aligned}
\mathbb{E}\left[\|g^k - \nabla f(x^*)\|_2^2 \mid x^k\right] \leq & 2\mathbb{E}\left[\|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)\|_2^2 \mid x^k\right] \\
& + 2\mathbb{E}\left[\|\nabla f_{i_k}(x^*) - y_{i_k}^k + \frac{1}{n}\sum_{j=1}^n y_j^k - \nabla f(x^*)\|_2^2 \mid x^k\right] \\
\leq & 2\mathbb{E}\left[\|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)\|_2^2 \mid x^k\right] \\
& + 2\mathbb{E}\left[\|\nabla f_{i_k}(x^*) - y_{i_k}^k\|_2^2 \mid x^k\right]
\end{aligned}
$$

SAGA: a proof

- We take mat.expectation, use smoothness (with convexity):

$$
\begin{aligned}
\mathbb{E}\left[\|g^k - \nabla f(x^*)\|_2^2 \mid x^k\right] \leq & 2\mathbb{E}\left[\|\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^*)\|_2^2 \mid x^k\right] \\
& + 2\mathbb{E}\left[\|\nabla f_{i_k}(x^*) - y_{i_k}^k\|_2^2 \mid x^k\right] \\
\leq & 4L \cdot \frac{1}{n}\sum_{i=1}^n (f_i(x^k) - f_i(x^*) - \langle \nabla f_i(x^k), x^k - x^* \rangle) \\
& + 2 \cdot \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2 \\
= & 4L \cdot (f(x^k) - f(x^*)) \\
& + 2 \cdot \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2
\end{aligned}
$$

SAGA: a proof

- Summarizing obtained results:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] = \|x^k - x^*\|_2^2 - 2\gamma\langle \mathbb{E}\left[g^k \mid x^k\right], x^k - x^*\rangle + \gamma^2 \mathbb{E}\left[\|g^k - \nabla f(x^*)\|_2^2 \mid x^k\right].$$

$$\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$$

$$\mathbb{E}\left[\|g^k - \nabla f(x^*)\|_2^2 \mid x^k\right] \leq 4L \cdot (f(x^k) - f(x^*)) + 2 \cdot \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2$$

SAGA: a proof

- Summarizing obtained results:

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2 \mid x^k\right] = \|x^k-x^*\|_2^2 - 2\gamma\langle \mathbb{E}\left[g^k \mid x^k\right], x^k-x^*\rangle + \gamma^2\mathbb{E}\left[\|g^k-\nabla f(x^*)\|_2^2 \mid x^k\right].$$

$$\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$$

$$\mathbb{E}\left[\|g^k-\nabla f(x^*)\|_2^2 \mid x^k\right] \leq 4L\cdot(f(x^k)-f(x^*)) + 2\cdot\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x^*)-y_i^k\|_2^2$$

- Putting it together:

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2 \mid x^k\right] \leq \|x^k-x^*\|_2^2 - 2\gamma\langle \nabla f(x^k), x^k-x^*\rangle + \gamma^2\left(4L\cdot(f(x^k)-f(x^*)) + 2\cdot\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x^*)-y_i^k\|_2^2\right)$$

SAGA: a proof

- Strong convexity of the function f:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] \leq (1 - \mu\gamma)\|x^k - x^*\|_2^2 - 2\gamma(1 - 2\gamma L)(f(x^k) - f(x^*)) + 2\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x^*) - y_i^k\|_2^2.$$

SAGA: a proof

- Strong convexity of the function f:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 \mid x^k\right] \leq (1-\mu\gamma)\|x^k - x^*\|_2^2 - 2\gamma(1-2\gamma L)(f(x^k) - f(x^*)) + 2\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*) - y_i^k\|_2^2.$$

- More formally, we came to the conclusion that if $y_i^k \to f_i(x^*)$, then the variance is «killed», and hence there will be linear convergence. Let us show how this can be strictly formalized.

SAGA: a proof

- Let's take a look at the behavior $\frac{1}{n}\sum\limits_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2$:

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2 \mid x^k\right] =& \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|y_i^{k+1} - \nabla f_i(x^*)\|_2^2 \mid x^k\right] \\
=& \left(1 - \frac{1}{n}\right)\cdot \frac{1}{n}\sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2 \\
&+ \frac{1}{n}\cdot\frac{1}{n}\sum_{i=1}^n \|f_i(x^k) - \nabla f_i(x^*)\|_2^2 \\
\leq& \left(1 - \frac{1}{n}\right)\cdot \frac{1}{n}\sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2 \\
&+ \frac{1}{n}\cdot 2L(f(x^k) - f(x^*)).
\end{aligned}
$$

SAGA: a proof

- Finally (here the full mathematical expectation is immediately thrown on):

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2\right] \leq (1-\mu\gamma)\mathbb{E}\left[\|x^k - x^*\|_2^2\right] - 2\gamma(1-2\gamma L)\mathbb{E}\left[f(x^k) - f(x^*)\right] + 2\gamma^2 \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2\right]$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2\right] \leq \left(1 - \frac{1}{n}\right) \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2\right] + \frac{1}{n} \cdot 2L\mathbb{E}\left[f(x^k) - f(x^*)\right].$$

SAGA: a proof

- Finally (here the full mathematical expectation is immediately thrown on):

$$\mathbb{E}\left[\|x^{k+1}-x^*\|_2^2\right] \leq (1-\mu\gamma)\mathbb{E}\left[\|x^k-x^*\|_2^2\right] - 2\gamma(1-2\gamma L)\mathbb{E}\left[f(x^k)-f(x^*)\right] + 2\gamma^2 \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^*) - y_i^k\|_2^2\right]$$

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2\right] \leq \left(1-\frac{1}{n}\right) \cdot \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2\right] + \frac{1}{n} \cdot 2L\mathbb{E}\left[f(x^k)-f(x^*)\right].$$

- We have two "converging" sequences, what remains is to neatly "concut" them together.

SAGA: a proof

- Пусть $M > 0$:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 + M\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2\right]$$

$$\leq (1-\mu\gamma)\mathbb{E}\left[\|x^k - x^*\|_2^2\right]$$

$$+ \left(1 + \frac{2}{M} - \frac{1}{n}\right)\mathbb{E}\left[M\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2\right]$$

$$- 2\gamma\left(1 - 2\gamma L - \frac{\gamma M L}{n}\right)\mathbb{E}\left[f(x^k) - f(x^*)\right]$$

SAGA: a proof

- Возьмем $M = 4n$:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^{n}\|y_i^{k+1} - \nabla f_i(x^*)\|_2^2\right]$$
$$\leq (1-\mu\gamma)\mathbb{E}\left[\|x^k - x^*\|_2^2\right]$$
$$+ \left(1 - \frac{1}{2n}\right)\mathbb{E}\left[4n\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^{n}\|y_i^{k} - \nabla f_i(x^*)\|_2^2\right]$$
$$- 2\gamma\left(1 - 6\gamma L\right)\mathbb{E}\left[f(x^k) - f(x^*)\right]$$

SAGA: a proof

- Возьмем $M = 4n$:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2\right]$$

$$\leq (1 - \mu\gamma)\mathbb{E}\left[\|x^k - x^*\|_2^2\right]$$

$$+ \left(1 - \frac{1}{2n}\right)\mathbb{E}\left[4n\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2\right]$$

$$- 2\gamma\left(1 - 6\gamma L\right)\mathbb{E}\left[f(x^k) - f(x^*)\right]$$

- Теперь $\gamma \leq \frac{1}{6L}$:

$$\mathbb{E}\left[\|x^{k+1} - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^n \|y_i^{k+1} - \nabla f_i(x^*)\|_2^2\right]$$

$$\leq \max\left\{(1 - \mu\gamma); \left(1 - \frac{1}{2n}\right)\right\}\mathbb{E}\left[\|x^k - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n}\sum_{i=1}^n \|y_i^k - \nabla f_i(\right.$$

SAGA: convergence

- We obtained convergence, but by an unusual criterion. The essence of the criterion is to reflect the physics of both convergence of $x^k \to x^*$ and $y_i^k \to \nabla f_i(x^*)$, which was put into the method.

Theorem (convergence of SAGA)

Let the unconstrained stochastic optimization problem of finite sum type with L-smooth, convex functions f_i and μ-strongly convex objective function f be solved by SAGA with $\gamma \leq \frac{1}{6L}$. Then the following convergence estimate is valid

$$\mathbb{E}\left[V_k\right] \leq \max\left\{(1-\mu\gamma); \left(1-\frac{1}{2n}\right)\right\}^k \mathbb{E}\left[V_0\right],$$

where $V_k = \|x^k - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n}\sum\limits_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2$.

SAGA: convergence

- We obtained convergence, but by an unusual criterion. The essence of the criterion is to reflect the physics of both convergence of $x^k \to x^*$ and $y_i^k \to \nabla f_i(x^*)$, which was put into the method.

Theorem (convergence of SAGA)

Let the unconstrained stochastic optimization problem of finite sum type with L-smooth, convex functions f_i and μ-strongly convex objective function f be solved by SAGA with $\gamma \leq \frac{1}{6L}$. Then the following convergence estimate is valid

$$\mathbb{E}\left[V_k\right] \leq \max\left\{(1-\mu\gamma); \left(1 - \frac{1}{2n}\right)\right\}^k \mathbb{E}\left[V_0\right],$$

where $V_k = \|x^k - x^*\|_2^2 + 4n\gamma^2 \cdot \frac{1}{n}\sum\limits_{i=1}^n \|y_i^k - \nabla f_i(x^*)\|_2^2$.

- It is easy to see that the convergence on $\mathbb{E}[V_k]$ also implies the convergence on $\mathbb{E}[\|x^k - x^*\|_2^2]$: $\mathbb{E}[\|x^k - x^*\|_2^2] \leq \mathbb{E}[V_k]$.

SVRG

Algorithm 3 SVRG

Input: step $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, number of iterations in epoch K, number of epochs S

for $s = 0, 1, \ldots, S-1$ **do**
  Update $w^s = x^{s-1,K}$
  Compute and save $\nabla f(w^s)$
  for $k = 0, 1, \ldots, K-1$ **do**
    $x^{s,k+1} = x^{s,k} - \gamma g^k$
    Generate i_k
    Compute $g^{k+1} = \nabla f_{i_k}(x^{s,k+1}) - \nabla f_{i_k}(w^s) + \nabla f(w^s)$
  end for
end for

Output: $x^{S-1,K}$

SVRG

- The idea – rarely count the full gradient at some reference point!

SVRG

- The idea – rarely count the full gradient at some reference point!
- $\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$.

SVRG

- The idea – rarely count the full gradient at some reference point!
- $\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$.
- If $x^k \to x^*$, we have $w^k \to x^*$, $(\nabla f_{i_k}(x^k) - \nabla f_{i_k}(w^k)) \to 0$, и $\nabla f(w^*) \to \nabla f(x^*) = 0$. Then $g^k \to 0$.

SVRG

- The idea – rarely count the full gradient at some reference point!
- $\mathbb{E}\left[g^k \mid x^k\right] = \nabla f(x^k)$.
- If $x^k \to x^*$, we have $w^k \to x^*$, $(\nabla f_{i_k}(x^k) - \nabla f_{i_k}(w^k)) \to 0$, и $\nabla f(w^*) \to \nabla f(x^*) = 0$. Then $g^k \to 0$.
- On the downside: you have to count the full gradient sometimes and calculate ∇f_{i_k} twice every iteration.

SARAH

Algorithm 4 SARAH

Input: step $\gamma > 0$, starting point $x^0 \in \mathbb{R}^d$, number of iterations in epoch K, number of epochs S

1: **for** $s = 0, 1, \ldots, S-1$ **do**
2: Compute $g^0 = \nabla f(x^{s-1,K})$
3: **for** $k = 0, 1, \ldots, K-1$ **do**
4: $x^{s,k+1} = x^{s,k} - \gamma g^k$
5: Generate independently i_k
6: Compute $g^{k+1} = \nabla f_{i_k}(x^{s,k-1}) - \nabla f_{i_k}(x^{s,k}) + g^k$
7: **end for**
8: **end for**

Output: $x^{S-1,K}$

SARAH

- The idea – to read the reference gradient more «smoothly» compared to SVRG!

SARAH

- The idea – to read the reference gradient more «smoothly» compared to SVRG!
- $\mathbb{E}\left[g^k \mid x^k\right] \neq \nabla f(x^k)$, but $\mathbb{E}\left[g^k\right] = \nabla f(x^k)$

SARAH

- The idea – to read the reference gradient more «smoothly» compared to SVRG!
- $\mathbb{E}\left[g^k \mid x^k\right] \neq \nabla f(x^k)$, but $\mathbb{E}\left[g^k\right] = \nabla f(x^k)$
- If $x^k \to x^*$, then $(\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^{k-1})) \to 0$, и $g^k \to$ const within the same epoch (launch of the internal cycle), but by virtue of the renewal of the $g^k = \nabla f(x^{s-1,K})$: $g^k \to 0$.

SARAH

- The idea – to read the reference gradient more «smoothly» compared to SVRG!
- $\mathbb{E}\left[g^k \mid x^k\right] \neq \nabla f(x^k)$, but $\mathbb{E}\left[g^k\right] = \nabla f(x^k)$
- If $x^k \to x^*$, then $(\nabla f_{i_k}(x^k) - \nabla f_{i_k}(x^{k-1})) \to 0$, и $g^k \to$ const within the same epoch (launch of the internal cycle), but by virtue of the renewal of the $g^k = \nabla f(x^{s-1,K})$: $g^k \to 0$.
- On the downside: you have to sometimes read the full gradient and each iteration has to be calculated twice ∇f_{i_k}.

Methods of variance reduction: bottom line

- Designed for stochastic problems of finite sum type (offline empirical risk minimization).

Methods of variance reduction: bottom line

- Designed for stochastic problems of finite sum type (offline empirical risk minimization).
- Provide convergence like that of gradient descent,

$$\text{Totally} \quad \mathcal{O}\left(\left[n + \frac{L}{\mu}\right] \log \frac{1}{\varepsilon}\right) \text{iterations for SAGA/SVRG/SARAH.}$$

but n times cheaper (we do not consider the full gradient, but only 1 summand).

Methods of variance reduction: bottom line

- Designed for stochastic problems of finite sum type (offline empirical risk minimization).
- Provide convergence like that of gradient descent,

$$\text{Totally} \quad \mathcal{O}\left(\left[n + \frac{L}{\mu}\right] \log \frac{1}{\varepsilon}\right) \text{ iterations for SAGA/SVRG/SARAH.}$$

  but n times cheaper (we do not consider the full gradient, but only 1 summand).
- Have disadvantages: wastes memory, counts full gradient.

Methods of variance reduction: bottom line

- Designed for stochastic problems of finite sum type (offline empirical risk minimization).
- Provide convergence like that of gradient descent,

$$\text{Totally} \quad \mathcal{O}\left(\left[n + \frac{L}{\mu}\right] \log \frac{1}{\varepsilon}\right) \text{ iterations for SAGA/SVRG/SARAH.}$$

  but n times cheaper (we do not consider the full gradient, but only 1 summand).
- Have disadvantages: wastes memory, counts full gradient.
- Can be accelerated (SVRG $\rightarrow$ Katyusha).