Stochastic Gradient Descent. Finite-sum problems.

Daniil Merkulov

Optimization for ML. Faculty of Computer Science. HSE University

Finite-sum problem

We consider classic finite-sample average minimization:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

The gradient descent acts like follows:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(x) \tag{GD}$$

- Convergence with constant α or line search.

Finite-sum problem

We consider classic finite-sample average minimization:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

The gradient descent acts like follows:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(x) \tag{GD}$$

- Convergence with constant α or line search.
- Iteration cost is linear in n. For ImageNet $n \approx 1.4 \cdot 10^7$, for WikiText $n \approx 10^8$.

Finite-sum problem

We consider classic finite-sample average minimization:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

The gradient descent acts like follows:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(x) \tag{GD}$$

- Convergence with constant α or line search.
- Iteration cost is linear in n. For ImageNet $n \approx 1.4 \cdot 10^7$, for WikiText $n \approx 10^8$.

Finite-sum problem

We consider classic finite-sample average minimization:

$$\min_{x \in \mathbb{R}^p} f(x) = \min_{x \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n f_i(x)$$

The gradient descent acts like follows:

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^n \nabla f_i(x) \tag{GD}$$

- Convergence with constant α or line search.
- Iteration cost is linear in n. For ImageNet $n \approx 1.4 \cdot 10^7$, for WikiText $n \approx 10^8$.

Let's/ switch from the full gradient calculation to its unbiased estimator, when we randomly choose i_k index of point at each iteration uniformly:

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k) \tag{SGD}$$

With $p(i_k = i) = \frac{1}{n}$, the stochastic gradient is an unbiased estimate of the gradient, given by:

$$\mathbb{E}[\nabla f_{i_k}(x)] = \sum_{i=1}^n p(i_k = i) \nabla f_i(x) = \sum_{i=1}^n \frac{1}{n} \nabla f_i(x) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x) = \nabla f(x)$$

This indicates that the expected value of the stochastic gradient is equal to the actual gradient of $f(x)$.

Results for Gradient Descent

Stochastic iterations are n times faster, but how many iterations are needed?

If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
Convex	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
Non-Convex	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$

- Stochastic has low iteration cost but slow convergence rate.

Results for Gradient Descent

Stochastic iterations are n times faster, but how many iterations are needed?

If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
Convex	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
Non-Convex	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$

- Stochastic has low iteration cost but slow convergence rate.
    - Sublinear rate even in strongly-convex case.

Results for Gradient Descent

Stochastic iterations are n times faster, but how many iterations are needed?

If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
Convex	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
Non-Convex	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$

- Stochastic has low iteration cost but slow convergence rate.
  - Sublinear rate even in strongly-convex case.
  - Bounds are unimprovable under standard assumptions.

Results for Gradient Descent

Stochastic iterations are n times faster, but how many iterations are needed?

If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
Convex	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
Non-Convex	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$

- Stochastic has low iteration cost but slow convergence rate.
    - Sublinear rate even in strongly-convex case.
    - Bounds are unimprovable under standard assumptions.
    - Oracle returns an unbiased gradient approximation with bounded variance.

Results for Gradient Descent

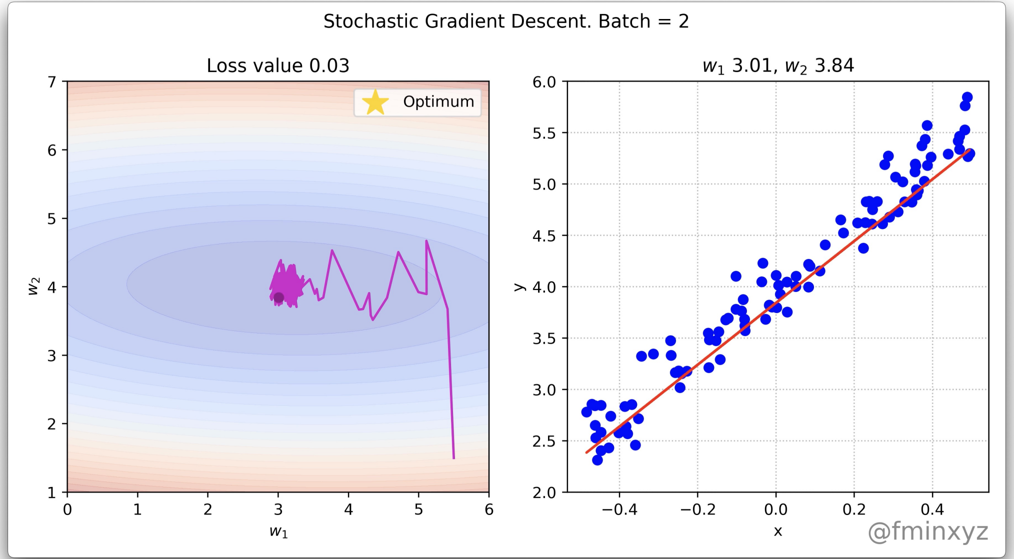
Stochastic iterations are n times faster, but how many iterations are needed?

If ∇f is Lipschitz continuous then we have:

Assumption	Deterministic Gradient Descent	Stochastic Gradient Descent
PL	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
Convex	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
Non-Convex	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$

- Stochastic has low iteration cost but slow convergence rate.
  - Sublinear rate even in strongly-convex case.
  - Bounds are unimprovable under standard assumptions.
  - Oracle returns an unbiased gradient approximation with bounded variance.
- Momentum and Quasi-Newton-like methods do not improve rates in stochastic case. Can only improve constant factors (bottleneck is variance, not condition number).

Typical behaviour



Convergence

Lipschitz continiity implies:

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

Convergence

Lipschitz continiity implies:

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

using (SGD):

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2} \|\nabla f_{i_k}(x_k)\|^2$$

Convergence

Lipschitz continiity implies:

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

using (SGD):

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2} \|\nabla f_{i_k}(x_k)\|^2$$

Now let's take expectation with respect to i_k:

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2} \|\nabla f_{i_k}(x_k)\|^2]$$

Convergence

Lipschitz continiity implies:

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

using (SGD):

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2}\|\nabla f_{i_k}(x_k)\|^2$$

Now let's take expectation with respect to i_k:

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2}\|\nabla f_{i_k}(x_k)\|^2]$$

Using linearity of expectation:

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \mathbb{E}[\nabla f_{i_k}(x_k)] \rangle + \alpha_k^2 \frac{L}{2}\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Convergence

Lipschitz continiity implies:

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

using (SGD):

$$f(x_{k+1}) \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2} \|\nabla f_{i_k}(x_k)\|^2$$

Now let's take expectation with respect to i_k:

$$\mathbb{E}[f(x_{k+1})] \leq \mathbb{E}[f(x_k) - \alpha_k \langle \nabla f(x_k), \nabla f_{i_k}(x_k) \rangle + \alpha_k^2 \frac{L}{2} \|\nabla f_{i_k}(x_k)\|^2]$$

Using linearity of expectation:

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \langle \nabla f(x_k), \mathbb{E}[\nabla f_{i_k}(x_k)] \rangle + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Since uniform sampling implies unbiased estimate of gradient: $\mathbb{E}[\nabla f_{i_k}(x_k)] = \nabla f(x_k)$:

$$\mathbb{E}[f(x_{k+1})] \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

Convergence. Smooth PL case.

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu(f(x) - f^*), \forall x \in \mathbb{R}^p \tag{PL}$$

Convergence. Smooth PL case.

$$
\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu(f(x) - f^*), \forall x \in \mathbb{R}^p \tag{PL}
$$

This inequality simply requires that the gradient grows faster than a quadratic function as we move away from the optimal function value. Note, that strong convexity implies PL, but not vice versa. Using PL we can write:

$$
\mathbb{E}[f(x_{k+1})] - f^* \leq (1 - 2\alpha_k \mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2}\mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]
$$

Convergence. Smooth PL case.

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu(f(x) - f^*), \forall x \in \mathbb{R}^p \tag{PL}$$

This inequality simply requires that the gradient grows faster than a quadratic function as we move away from the optimal function value. Note, that strong convexity implies PL, but not vice versa. Using PL we can write:

$$\mathbb{E}[f(x_{k+1})] - f^* \leq (1 - 2\alpha_k \mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

This bound already indicates, that we have something like linear convergence if far from solution and gradients are similar, but no progress if close to solution or have high variance in gradients at the same time.

Convergence. Smooth PL case.

$$
\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu(f(x) - f^*), \forall x \in \mathbb{R}^p \tag{PL}
$$

This inequality simply requires that the gradient grows faster than a quadratic function as we move away from the optimal function value. Note, that strong convexity implies PL, but not vice versa. Using PL we can write:

$$
\mathbb{E}[f(x_{k+1})] - f^* \leq (1 - 2\alpha_k \mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]
$$

This bound already indicates, that we have something like linear convergence if far from solution and gradients are similar, but no progress if close to solution or have high variance in gradients at the same time.

Now we assume, that the variance of the stochastic gradients is bounded:

$$
\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2
$$

Convergence. Smooth PL case.

$$\frac{1}{2}\|\nabla f(x)\|_2^2 \geq \mu(f(x) - f^*), \forall x \in \mathbb{R}^p \tag{PL}$$

This inequality simply requires that the gradient grows faster than a quadratic function as we move away from the optimal function value. Note, that strong convexity implies PL, but not vice versa. Using PL we can write:

$$\mathbb{E}[f(x_{k+1})] - f^* \leq (1 - 2\alpha_k \mu)[f(x_k) - f^*] + \alpha_k^2 \frac{L}{2} \mathbb{E}[\|\nabla f_{i_k}(x_k)\|^2]$$

This bound already indicates, that we have something like linear convergence if far from solution and gradients are similar, but no progress if close to solution or have high variance in gradients at the same time.

Now we assume, that the variance of the stochastic gradients is bounded:

$$\mathbb{E}[\|\nabla f_i(x_k)\|^2] \leq \sigma^2$$

Thus, we have

$$\mathbb{E}[f(x_{k+1}) - f^*] \leq (1 - 2\alpha_k \mu)[f(x_k) - f^*] + \frac{L\sigma^2\alpha_k^2}{2}.$$

Convergence. Smooth PL case.

1. Consider **decreasing stepsize** strategy with $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ we obtain

$$\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2(2k+1)^2|}{8\mu^2(k+1)^4}$$

Convergence. Smooth PL case.

1. Consider **decreasing stepsize** strategy with $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ we obtain

$$
\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2(2k+1)^2|}{8\mu^2(k+1)^4}
$$

2. Multiplying both sides by $(k+1)^2$ and letting $\delta_f(k) \equiv k^2 \mathbb{E}[f(x_k) - f^*]$ we get

$$
\begin{aligned}
\delta_f(k+1) &\leq \delta_f(k) + \frac{L\sigma^2(2k+1)^2}{8\mu^2(k+1)^2} \\
&\leq \delta_f(k) + \frac{L\sigma^2}{2\mu^2},
\end{aligned}
$$

Convergence. Smooth PL case.

1. Consider **decreasing stepsize** strategy with $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ we obtain

$$
\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2(2k+1)^2|}{8\mu^2(k+1)^4}
$$

2. Multiplying both sides by $(k+1)^2$ and letting $\delta_f(k) \equiv k^2\mathbb{E}[f(x_k) - f^*]$ we get

$$
\begin{aligned}
\delta_f(k+1) &\leq \delta_f(k) + \frac{L\sigma^2(2k+1)^2}{8\mu^2(k+1)^2} \\
&\leq \delta_f(k) + \frac{L\sigma^2}{2\mu^2},
\end{aligned}
$$

Convergence. Smooth PL case.

1. Consider **decreasing stepsize** strategy with $\alpha_k = \frac{2k+1}{2\mu(k+1)^2}$ we obtain

$$
\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{k^2}{(k+1)^2}[f(x_k) - f^*] + \frac{L\sigma^2(2k+1)^2|}{8\mu^2(k+1)^4}
$$

2. Multiplying both sides by $(k+1)^2$ and letting $\delta_f(k) \equiv k^2\mathbb{E}[f(x_k) - f^*]$ we get

$$
\begin{aligned}
\delta_f(k+1) &\leq \delta_f(k) + \frac{L\sigma^2(2k+1)^2}{8\mu^2(k+1)^2} \\
&\leq \delta_f(k) + \frac{L\sigma^2}{2\mu^2},
\end{aligned}
$$

where the second line follows from $\frac{2k+1}{k+1} < 2$. Summing up this inequality from $k = 0$ to k and using the fact that $\delta_f(0) = 0$ we get

$$
\delta_f(k+1) \leq \delta_f(0) + \frac{L\sigma^2}{2\mu^2}\sum_{i=0}^{k} 1 \leq \frac{L\sigma^2(k+1)}{2\mu^2} \Rightarrow \quad (k+1)^2\mathbb{E}[f(x_{k+1}) - f^*] \leq \frac{L\sigma^2(k+1)}{2\mu^2}
$$

which gives the stated rate.

Convergence. Smooth PL case.

3. **Constant step size**: Choosing $\alpha_k = \alpha$ for any $\alpha < 1/2\mu$ yields

$$
\begin{aligned}
\mathbb{E}[f(x_{k+1}) - f^*] &\leq (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha^2}{2} \sum_{i=0}^{k} (1 - 2\alpha\mu)^i \\
&\leq (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha^2}{2} \sum_{i=0}^{\infty} (1 - 2\alpha\mu)^i \\
&= (1 - 2\alpha\mu)^k [f(x_0) - f^*] + \frac{L\sigma^2\alpha}{4\mu},
\end{aligned}
$$

where the last line uses that $\alpha < 1/2\mu$ and the limit of the geometric series.

Convergence. Smooth non-convex case.

Convergence. Convex case.

Mini-batch SGD

Approach 1: Control the sample size

The deterministic method uses all n gradients:

$$\nabla f(x_k) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k).$$

The stochastic method approximates this using just 1 sample:

$$\nabla f_{ik}(x_k) \approx \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k).$$

A common variant is to use a larger sample B_k ("mini-batch"):

$$\frac{1}{|B_k|}\sum_{i \in B_k} \nabla f_i(x_k) \approx \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_k),$$

particularly useful for vectorization and parallelization.
For example, with 16 cores set $|B_k| = 16$ and compute 16 gradients at once.

Mini-Batching as Gradient Descent with Error

The SG method with a sample B_k ("mini-batch") uses iterations:

$$x_{k+1} = x_k - \alpha_k \left(\frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(x_k) \right).$$

Let's view this as a "gradient method with error":

$$x_{k+1} = x_k - \alpha_k (\nabla f(x_k) + e_k),$$

where e_k is the difference between the approximate and true gradient.

If you use $\alpha_k = \frac{1}{L}$, then using the descent lemma, this algorithm has:

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

for any error e_k.

Effect of Error on Convergence Rate

Our progress bound with $\alpha_k = \frac{1}{L}$ and error in the gradient of e_k is:

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

Connection between "error-free" rate and "with error" rate:

- If the "error-free" rate is $O(\frac{1}{k})$, you maintain this rate if $\|e_k\|^2 = O(\frac{1}{k})$.

Effect of Error on Convergence Rate

Our progress bound with $\alpha_k = \frac{1}{L}$ and error in the gradient of e_k is:

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

Connection between "error-free" rate and "with error" rate:

- If the "error-free" rate is $O(\frac{1}{k})$, you maintain this rate if $\|e_k\|^2 = O(\frac{1}{k})$.
- If the "error-free" rate is $O(\rho^k)$, you maintain this rate if $\|e_k\|^2 = O(\rho^k)$.

Effect of Error on Convergence Rate

Our progress bound with $\alpha_k = \frac{1}{L}$ and error in the gradient of e_k is:

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

Connection between "error-free" rate and "with error" rate:

- If the "error-free" rate is $O(\frac{1}{k})$, you maintain this rate if $\|e_k\|^2 = O(\frac{1}{k})$.
- If the "error-free" rate is $O(\rho^k)$, you maintain this rate if $\|e_k\|^2 = O(\rho^k)$.

Effect of Error on Convergence Rate

Our progress bound with $\alpha_k = \frac{1}{L}$ and error in the gradient of e_k is:

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

Connection between "error-free" rate and "with error" rate:

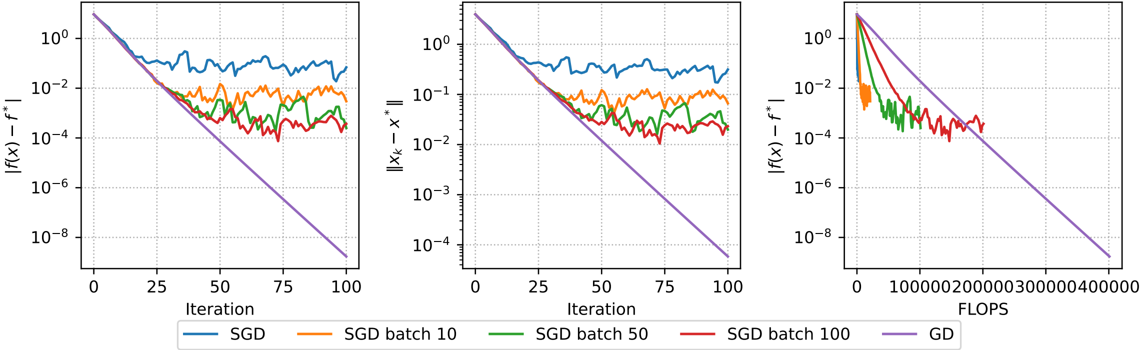
- If the "error-free" rate is $O(\frac{1}{k})$, you maintain this rate if $\|e_k\|^2 = O(\frac{1}{k})$.
- If the "error-free" rate is $O(\rho^k)$, you maintain this rate if $\|e_k\|^2 = O(\rho^k)$.

If the error goes to zero more slowly, then the rate at which it goes to zero becomes the bottleneck.
So, to understand the effect of batch size, we need to know how $|B_k|$ affects $\|e_k\|^2$.

Main problem of SGD

$$
f(x) = \frac{\mu}{2}\|x\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\log(1+\exp(-y_i\langle a_i, x\rangle)) \to \min_{x\in\mathbb{R}^n}
$$

Strongly convex binary logistic regression. m=200, n=10, mu=1.



Conclusions

- SGD with fixed learning rate does not converge even for PL (strongly convex) case

Conclusions

- SGD with fixed learning rate does not converge even for PL (strongly convex) case
- SGD achieves sublinear convergence with rate $\mathcal{O}\left(\frac{1}{k}\right)$ for PL-case.

Conclusions

- SGD with fixed learning rate does not converge even for PL (strongly convex) case
- SGD achieves sublinear convergence with rate $\mathcal{O}\left(\frac{1}{k}\right)$ for PL-case.
- Nesterov/Polyak accelerations do not improve convergence rate

Conclusions

- SGD with fixed learning rate does not converge even for PL (strongly convex) case
- SGD achieves sublinear convergence with rate $\mathcal{O}\left(\frac{1}{k}\right)$ for PL-case.
- Nesterov/Polyak accelerations do not improve convergence rate
- Two-phase Newton-like method achieves $\mathcal{O}\left(\frac{1}{k}\right)$ without strong convexity.