



CONVERGENCE RATES FOR OPTIMIZATION ALGORITHMS

SOURANGSHU BHATTACHARYA

CSE, IIT KHARAGPUR

WEB: [HTTPS://CSE.IITKGP.AC.IN/~SOURANGSHU/](https://cse.iitkgp.ac.in/~sourangshu/)

EMAIL: SOURANGSHU@CSE.IITKGP.AC.IN



CONVERGENCE RATES



CONVERGENCE RATE AND ASSUMPTIONS

A sequence $\{x^k\}$ is said to converge at the rate γ^k , if:

$$\|\mathbf{x}^{k+1} - \mathbf{x}^*\| \leq \gamma \|\mathbf{x}^k - \mathbf{x}^*\| \quad (\Rightarrow \|\mathbf{x}^k - \mathbf{x}^*\| \leq \gamma^k \|\mathbf{x}^0 - \mathbf{x}^*\|),$$

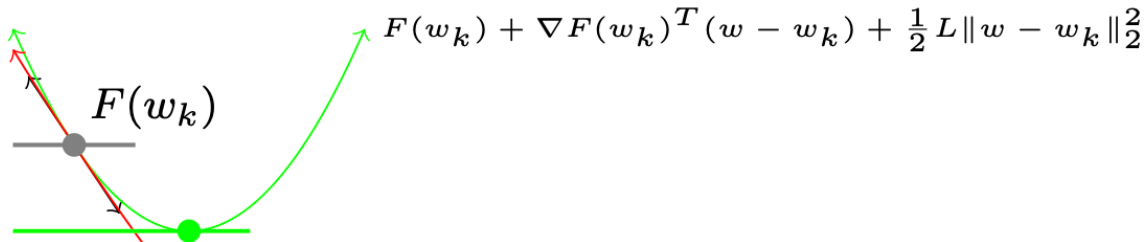
Assumption $\langle L/c \rangle$

The objective function $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is

- ▶ c -strongly convex (\Rightarrow unique minimizer) and
- ▶ L -smooth (i.e., ∇F is Lipschitz continuous with constant L).



GRADIENT DESCENT CONVERGENCE



$$F(w_k) + \nabla F(w_k)^T (w - w_k) + \frac{1}{2} c \|w - w_k\|_2^2$$

Choosing $\alpha = 1/L$ to minimize upper bound yields

$$(F(w_{k+1}) - F_*) \leq (F(w_k) - F_*) - \frac{1}{2L} \|\nabla F(w_k)\|_2^2$$

while lower bound yields

$$\frac{1}{2} \|\nabla F(w_k)\|_2^2 \geq c(F(w_k) - F_*),$$

which together imply that

$$(F(w_{k+1}) - F_*) \leq (1 - \frac{c}{L})(F(w_k) - F_*).$$

w_k

w



L-SMOOTH OBJECTIVE FUNCTION

$$\|\nabla F(w) - \nabla F(\bar{w})\| \leq L\|w - \bar{w}\|$$

Proof of inequality:

$$\begin{aligned} F(w) &= F(\bar{w}) + \int_0^1 \frac{\partial F(\bar{w} + t(w - \bar{w}))}{\partial t} dt \\ &= F(\bar{w}) + \int_0^1 \nabla F(\bar{w} + t(w - \bar{w}))^T (w - \bar{w}) dt \\ &= F(\bar{w}) + \nabla F(\bar{w})^T (w - \bar{w}) + \int_0^1 [\nabla F(\bar{w} + t(w - \bar{w})) - \nabla F(\bar{w})]^T (w - \bar{w}) dt \\ &\leq F(\bar{w}) + \nabla F(\bar{w})^T (w - \bar{w}) + \int_0^1 L\|t(w - \bar{w})\|_2 \|w - \bar{w}\|_2 dt, \end{aligned}$$



C-STRONG CONVEXITY

$$F(w) \geq F(w_k) + \nabla F(w_k)^T (w - w_k) + \frac{c}{2} \|w - w_k\|^2$$

Minimizing the RHS w.r.t. w :

$$\tilde{w} = w_k - \frac{1}{c} \nabla F(w_k)$$

Lower bound on RHS: $F(w_k) - \frac{1}{2c} \|\nabla F(w_k)\|^2$

Putting back in the first equation:

$$c(F(w_k) - F(w)) \leq 1/2 \|\nabla F(w_k)\|^2$$



Convergence Rate and Computational Complexity

Overall Complexity (ϵ) = Convergence Rate⁻¹(ϵ) * Complexity of each iteration

	Strongly Convex + Smooth			Convex + Smooth		
	Convergence Rate	Complexity of each iteration	Overall Complexity	Convergence Rate	Complexity of each iteration	Overall Complexity
GD	$O\left(\exp\left(-\frac{t}{Q}\right)\right)$	$O(n \cdot d)$	$O\left(nd \cdot Q \cdot \log\left(\frac{1}{\epsilon}\right)\right)$	$O\left(\frac{\beta}{t}\right)$	$O(n \cdot d)$	$O\left(nd \cdot \beta \cdot \left(\frac{1}{\epsilon}\right)\right)$
SGD	$O\left(\frac{1}{t}\right)$	$O(d)$	$O\left(\frac{d}{\epsilon}\right)$	$O\left(\frac{1}{\sqrt{t}}\right)$	$O(d)$	$O\left(\frac{d}{\epsilon^2}\right)$



SGD ANALYSIS

THEOREM 14.8 *Let $B, \rho > 0$. Let f be a convex function and let $\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w}: \|\mathbf{w}\| \leq B} f(\mathbf{w})$. Assume that SGD is run for T iterations with $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$. Assume also that for all t , $\|\mathbf{v}_t\| \leq \rho$ with probability 1. Then,*

$$\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^*) \leq \frac{B\rho}{\sqrt{T}}.$$

Therefore, for any $\epsilon > 0$, to achieve $\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^) \leq \epsilon$, it suffices to run the SGD algorithm for a number of iterations that satisfies*

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}.$$



SGD ANALYSIS

LEMMA 14.1 *Let $\mathbf{v}_1, \dots, \mathbf{v}_T$ be an arbitrary sequence of vectors. Any algorithm with an initialization $\mathbf{w}^{(1)} = 0$ and an update rule of the form*

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta \mathbf{v}_t \quad (14.4)$$

satisfies

$$\sum_{t=1}^T \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle \leq \frac{\|\mathbf{w}^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|\mathbf{v}_t\|^2. \quad (14.5)$$

In particular, for every $B, \rho > 0$, if for all t we have that $\|\mathbf{v}_t\| \leq \rho$ and if we set $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, then for every \mathbf{w}^ with $\|\mathbf{w}^*\| \leq B$ we have*

$$\frac{1}{T} \sum_{t=1}^T \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle \leq \frac{B\rho}{\sqrt{T}}.$$



SGD ANALYSIS

Proof Using algebraic manipulations (completing the square), we obtain:

$$\begin{aligned}\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle &= \frac{1}{\eta} \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \eta \mathbf{v}_t \rangle \\ &= \frac{1}{2\eta} (-\|\mathbf{w}^{(t)} - \mathbf{w}^* - \eta \mathbf{v}_t\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 + \eta^2 \|\mathbf{v}_t\|^2) \\ &= \frac{1}{2\eta} (-\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2) + \frac{\eta}{2} \|\mathbf{v}_t\|^2,\end{aligned}$$



SGD ANALYSIS

where the last equality follows from the definition of the update rule. Summing the equality over t , we have

$$\sum_{t=1}^T \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle = \frac{1}{2\eta} \sum_{t=1}^T \left(-\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|^2 + \|\mathbf{w}^{(t)} - \mathbf{w}^*\|^2 \right) + \frac{\eta}{2} \sum_{t=1}^T \|\mathbf{v}_t\|^2. \quad (14.6)$$

The first sum on the right-hand side is a telescopic sum that collapses to

$$\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2.$$



SGD ANALYSIS

Plugging this in Equation (14.6), we have

$$\begin{aligned}\sum_{t=1}^T \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle &= \frac{1}{2\eta} (\|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 - \|\mathbf{w}^{(T+1)} - \mathbf{w}^*\|^2) + \frac{\eta}{2} \sum_{t=1}^T \|\mathbf{v}_t\|^2 \\ &\leq \frac{1}{2\eta} \|\mathbf{w}^{(1)} - \mathbf{w}^*\|^2 + \frac{\eta}{2} \sum_{t=1}^T \|\mathbf{v}_t\|^2 \\ &= \frac{1}{2\eta} \|\mathbf{w}^*\|^2 + \frac{\eta}{2} \sum_{t=1}^T \|\mathbf{v}_t\|^2,\end{aligned}$$

where the last equality is due to the definition $\mathbf{w}^{(1)} = 0$. This proves the first part of the lemma (Equation (14.5)). The second part follows by upper bounding $\|\mathbf{w}^*\|$ by B , $\|\mathbf{v}_t\|$ by ρ , dividing by T , and plugging in the value of η . \square



SGD ANALYSIS

$$\mathbb{E}_{\mathbf{v}_{1:T}} [f(\bar{\mathbf{w}}) - f(\mathbf{w}^*)] \leq \mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T (f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)) \right].$$

Since Lemma 14.1 holds for any sequence $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_T$, it applies to SGD as well. By taking expectation of the bound in the lemma we have

$$\mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle \right] \leq \frac{B\rho}{\sqrt{T}}. \quad (14.9)$$

It is left to show that

$$\mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T (f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)) \right] \leq \mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle \right], \quad (14.10)$$



SGD ANALYSIS

Using the linearity of the expectation we have

$$\mathbb{E}_{\mathbf{v}_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle \right] = \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{\mathbf{v}_{1:T}} [\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle].$$

Next, we recall the *law of total expectation*: For every two random variables α, β , and a function g , $\mathbb{E}_{\alpha}[g(\alpha)] = \mathbb{E}_{\beta} \mathbb{E}_{\alpha}[g(\alpha)|\beta]$. Setting $\alpha = \mathbf{v}_{1:t}$ and $\beta = \mathbf{v}_{1:t-1}$ we get that

$$\begin{aligned} \mathbb{E}_{\mathbf{v}_{1:T}} [\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle] &= \mathbb{E}_{\mathbf{v}_{1:t}} [\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle] \\ &= \mathbb{E}_{\mathbf{v}_{1:t-1}} \mathbb{E}_{\mathbf{v}_{1:t}} [\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle | \mathbf{v}_{1:t-1}]. \end{aligned}$$

Once we know $\mathbf{v}_{1:t-1}$, the value of $\mathbf{w}^{(t)}$ is not random any more and therefore

$$\mathbb{E}_{\mathbf{v}_{1:t-1}} \mathbb{E}_{\mathbf{v}_{1:t}} [\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle | \mathbf{v}_{1:t-1}] = \mathbb{E}_{\mathbf{v}_{1:t-1}} \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbb{E}_{\mathbf{v}_t} [\mathbf{v}_t | \mathbf{v}_{1:t-1}] \rangle.$$



SGD ANALYSIS

Since $\mathbf{w}^{(t)}$ only depends on $\mathbf{v}_{1:t-1}$ and SGD requires that $\mathbb{E}_{\mathbf{v}_t}[\mathbf{v}_t | \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$ we obtain that $\mathbb{E}_{\mathbf{v}_t}[\mathbf{v}_t | \mathbf{v}_{1:t-1}] \in \partial f(\mathbf{w}^{(t)})$. Thus,

$$\mathbb{E}_{\mathbf{v}_{1:t-1}} \langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbb{E}_{\mathbf{v}_t}[\mathbf{v}_t | \mathbf{v}_{1:t-1}] \rangle \geq \mathbb{E}_{\mathbf{v}_{1:t-1}} [f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)].$$

Overall, we have shown that

$$\begin{aligned} \mathbb{E}_{\mathbf{v}_{1:T}} [\langle \mathbf{w}^{(t)} - \mathbf{w}^*, \mathbf{v}_t \rangle] &\geq \mathbb{E}_{\mathbf{v}_{1:t-1}} [f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)] \\ &= \mathbb{E}_{\mathbf{v}_{1:T}} [f(\mathbf{w}^{(t)}) - f(\mathbf{w}^*)]. \end{aligned}$$

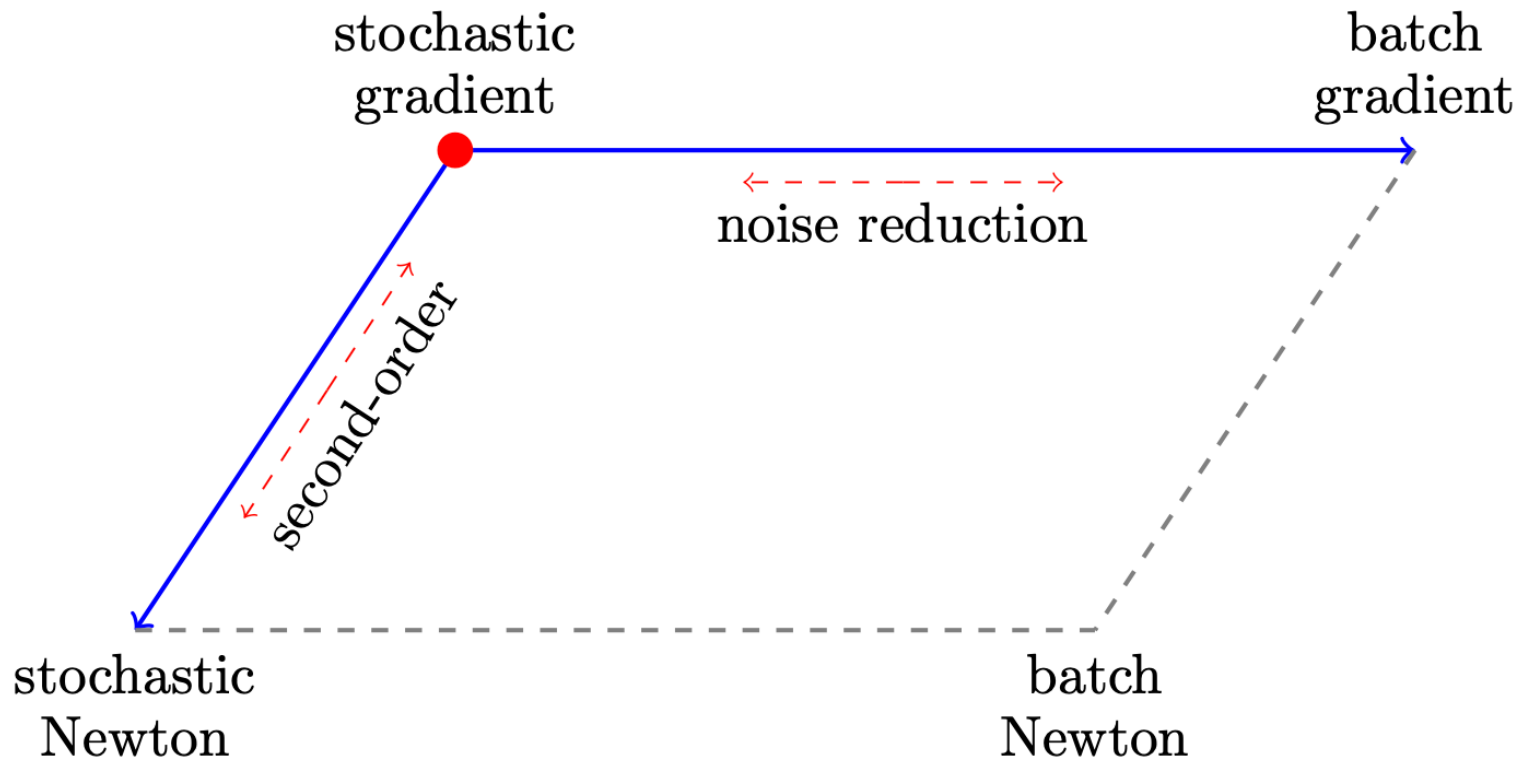
Summing over t , dividing by T , and using the linearity of expectation, we get that Equation (14.10) holds, which concludes our proof. \square



LINEAR RATE METHODS



IMPROVING SGD



Slides taken from Jorge Nocedal



STOCHASTIC AVERAGED GRADIENT

- Can we have a rate of $O(\rho^t)$ with only 1 gradient evaluation per iteration?
 - YES! The **stochastic average gradient (SAG)** algorithm:
 - Randomly select i_t from $\{1, 2, \dots, N\}$ and compute $f'_{i_t}(x^t)$.

$$x^{t+1} = x^t - \frac{\alpha^t}{N} \sum_{i=1}^N y_i^t$$

- **Memory:** $y_i^t = \nabla f_i(x^t)$ from the **last t** where i was selected.
[Le Roux et al., 2012]
 - **Stochastic** variant of increment average gradient (IAG).
[Blatt et al., 2007]
 - Assumes gradients of non-selected examples don't change.
 - Assumption becomes accurate as $\|x^{t+1} - x^t\| \rightarrow 0$.



SAG CONVERGENCE RATE

- If each f'_i is L -continuous and f is strongly-convex, with $\alpha_t = 1/16L$ SAG has

$$\mathbb{E}[f(x^t) - f(x^*)] \leq \left(1 - \min\left\{\frac{\mu}{16L}, \frac{1}{8N}\right\}\right)^t C,$$

where

$$C = [f(x^0) - f(x^*)] + \frac{4L}{N} \|x^0 - x^*\|^2 + \frac{\sigma^2}{16L}.$$

- Linear convergence rate but only 1 gradient per iteration.
 - For well-conditioned problems, constant reduction per pass:

$$\left(1 - \frac{1}{8N}\right)^N \leq \exp\left(-\frac{1}{8}\right) = 0.8825.$$

- For ill-conditioned problems, almost same as deterministic method (but N times faster).



SAG CONVERGENCE RATE

- Assume that $N = 700000$, $L = 0.25$, $\mu = 1/N$:
 - Gradient method has rate $\left(\frac{L-\mu}{L+\mu}\right)^2 = 0.99998$.
 - Accelerated gradient method has rate $(1 - \sqrt{\frac{\mu}{L}}) = 0.99761$.
 - **SAG (N iterations)** has rate $(1 - \min\{\frac{\mu}{16L}, \frac{1}{8N}\})^N = 0.88250$.
 - *Fastest possible* first-order method: $\left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)^2 = 0.99048$.
- **SAG beats two lower bounds:**
 - Stochastic gradient bound (of $O(1/t)$).
 - Deterministic gradient bound (for typical L , μ , and N).
- Number of f'_i evaluations to reach ϵ :
 - Stochastic: $O(\frac{L}{\mu}(1/\epsilon))$.
 - Gradient: $O(N\frac{L}{\mu}\log(1/\epsilon))$.
 - Accelerated: $O(N\sqrt{\frac{L}{\mu}}\log(1/\epsilon))$.
 - **SAG:** $O(\max\{N, \frac{L}{\mu}\}\log(1/\epsilon))$.



SAG IMPLEMENTATION

- Basic SAG algorithm:
 - while(1)
 - Sample i from $\{1, 2, \dots, N\}$.
 - Compute $f'_i(x)$.
 - $d = d - y_i + f'_i(x)$.
 - $y_i = f'_i(x)$.
 - $x = x - \frac{\alpha}{N}d$.
- Practical variants of the basic algorithm allow:
 - Regularization.
 - Sparse gradients.
 - Automatic step-size selection.
 - Common to use adaptive step-size procedure to estimate L .
 - Termination criterion.
 - Can use $\|x^{t+1} - x^t\|/\alpha = \frac{1}{n}d \approx \|\nabla f(x^t)\|$ to decide when to stop.
 - Acceleration [Lin et al., 2015].
 - Adaptive non-uniform sampling [Schmidt et al., 2013].



SAG IMPLEMENTATION

- Does **re-shuffling** and doing full passes work better?
 - For classic SG: **Maybe?**
 - Noncommutative arithmetic-geometric mean inequality conjecture.
 - For SAG: **NO.**
 - Performance is intermediate between IAG and SAG.
- Can **non-uniform** sampling help?
 - For classic SG methods, can only improve constants.
 - For SAG, **bias sampling towards Lipschitz constants L_i ,**

[Recht & Ré, 2012]

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|.$$

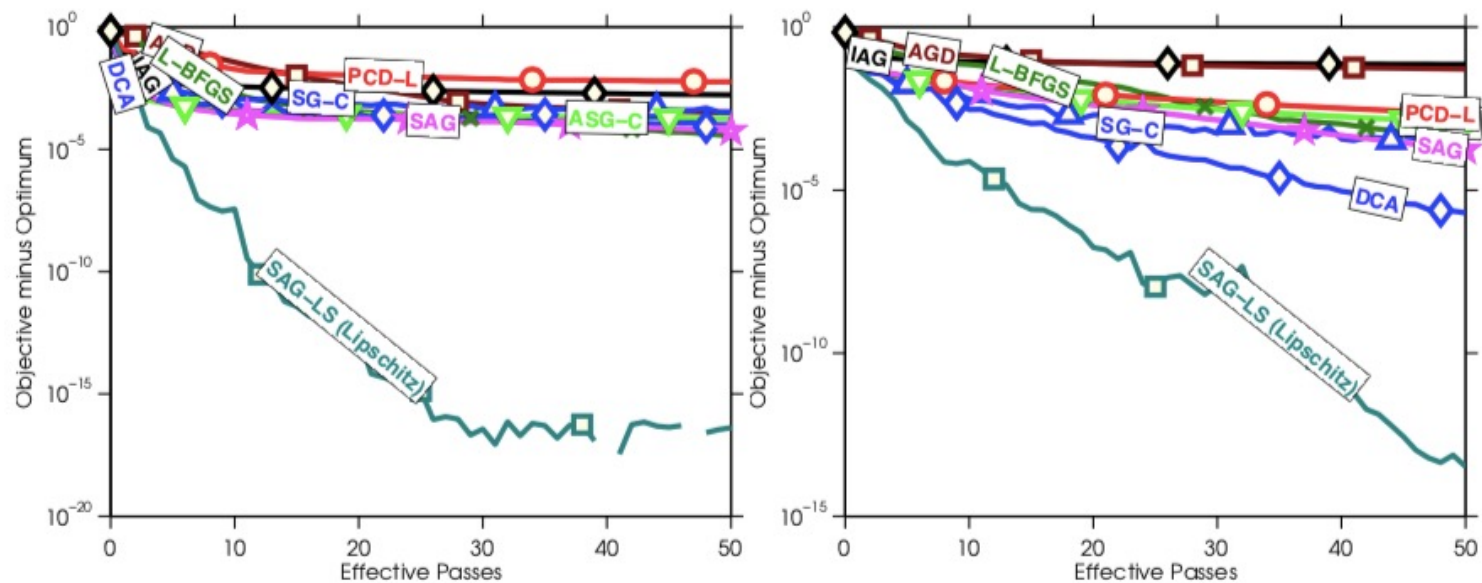
improves rate to depend on L_{mean} instead of L_{max} .

(with bigger step size)

- **Adaptively estimate L_i as you go.** (see paper/code).
- Slowly learns to **ignore well-classified examples.**

SAG with Non-Uniform Sampling

- protein ($n = 145751$, $p = 74$) and sido ($n = 12678$, $p = 4932$)



- Adaptive non-uniform sampling helps a lot.



STOCHASTIC VARIANCE REDUCED GD

SVRG algorithm:

- Start with x_0
- for $s = 0, 1, 2, \dots$
 - $d_s = \frac{1}{N} \sum_{i=1}^N f'_i(x_s)$
 - $x^0 = x_s$
 - for $t = 1, 2, \dots, m$
 - Randomly pick $i_t \in \{1, 2, \dots, N\}$
 - $x^t = x^{t-1} - \alpha_t (f'_{i_t}(x^{t-1}) - f'_{i_t}(x_s) + d_s)$.
 - $x_{s+1} = x^t$ for random $t \in \{1, 2, \dots, m\}$.

Requires **2 gradients per iteration and occasional full passes**, but **only requires storing d_s and x_s** .

Practical issues similar to SAG (acceleration versions, automatic step-size/termination, handles sparsity/regularization, non-uniform sampling, mini-batches).



References:

- Convergence rate analysis of GD and SGD:
Understanding Machine Learning: Theory to Algorithms
Shai Shalev Shwartz and Shai Ben David.



THANKS

QUESTIONS?

Email: sourangshu@cse.iitkgp.ac.in