Deep Learning

Lecture 3: Large-Scale Learning: Convex case

University of Agder,
Kristiansand, Norway

Prepared by: Hadi Ghauch⋆, Hossein S. Ghadikolaei†

⋆ Telecom ParisTech

† Royal Institute of Technology, KTH

https://sites.google.com/view/fundl/home

April 2019

Learning outcomes

- Recap of (deterministic) iterative algorithms for convex optimization
- Stochastic optimization
- Variance reduction techniques
- Convergence analysis

Outline

Outline

Deterministic Algorithms

Recap from Lecture 2

Smooth strongly convex problems (L-smooth, μ-strong convexity)
minimize$_{\boldsymbol{w}\in\mathbb{R}^d}$ $f(\boldsymbol{w})$

Gradient descent: move along steepest descent

Newton's methods: include Hessian (curvature) information in the update. Not used in learning

Nesterov Acceleration: mix between interpolation and gradient descent (need to know, L)

Momentum Methods (Heavy ball method): add momentum term to reduce the oscillations and improve convergence (need to know, L, μ)

Other deterministic algorithms (not covered in Lec 2)

Projected gradient descent: minimize$_{\boldsymbol{w}\in\mathcal{W}}$ $f(\boldsymbol{w})$
Nonsmooth strongly convex problems (μ-strong convexity)
Subgradient methods: minimize$_{\boldsymbol{w}\in\mathbb{R}^d}$ $f(\boldsymbol{w})$
Proximal methods: minimize$_{\boldsymbol{w}\in\mathbb{R}^d}$ $g(\boldsymbol{w}) + h(\boldsymbol{w})$
Accelerated proximal methods

Recap of Basic definitions

Convexity for differentiable function:

$$f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1) \geq \nabla f(\boldsymbol{w}_1)^T(\boldsymbol{w}_2 - \boldsymbol{w}_1)$$

Strongly convexity:

$$f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1) \geq \nabla f(\boldsymbol{w}_1)^T(\boldsymbol{w}_2 - \boldsymbol{w}_1) + \frac{\mu}{2}\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2^2$$

Smoothness:

$$f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1) \leq \nabla f(\boldsymbol{w}_1)^T(\boldsymbol{w}_2 - \boldsymbol{w}_1) + \frac{L}{2}\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2^2$$

Bounded error for initial guess: $\mathbb{E}\left[\|\boldsymbol{w}_1 - \boldsymbol{w}^\star\|_2\right] \leq R$

Lipschitz continuity (bounded gradients)

$$\|\boldsymbol{w}\|_2 \leq D \Rightarrow \|\nabla f(\boldsymbol{w})\|_2 \leq B$$

$$\text{or} \quad \|\boldsymbol{w}_1\|_2, \|\boldsymbol{w}_2\|_2 \leq D \Rightarrow |f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1)| \leq B\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|_2$$

Outline

Setting and Motivation

- **Batch GD:** Let $f(\boldsymbol{w}) = \frac{1}{N}\sum_{i\in[N]} f_i(\boldsymbol{w})$

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k) = \boldsymbol{w}_k - \alpha_k \left(\frac{1}{N} \sum_{i\in[N]} \nabla f_i(\boldsymbol{w}_k) \right)$$

  evaluate gradient of N samples: $O(N)$
  N 'large' for modern datasets: even $O(N)$ complexity is too high.
  idea: compute grad over subset of training samples

- **Stochastic gradient (SG) methods:**

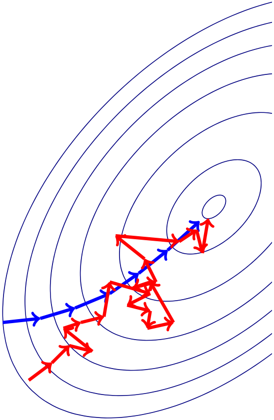
$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \; g(\boldsymbol{w}_k; \zeta_k) = \boldsymbol{w}_k - \alpha_k \widehat{\nabla} f(\boldsymbol{w}_k)$$

  ζ_k is a random subset of $[N]$
  $g(\boldsymbol{w}_k; \zeta_k)$ is a noisy version ("estimation") of $\nabla f(\boldsymbol{w}_k)$.

Method	**Per iteration cost**	**# iterations**
GD	Expensive (usually linear in N)	Usually few
SG	Very cheap, independent of N	Many

Main tradeoff: Per-iteration cost vs per-iteration improvement

Setting and Motivation



Considerations for SGD method

Good theoretical guarantees: Consider strongly convex smooth f, then

- GD: $f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star) \leq \mathcal{O}(\rho^k)$, $\rho \in (0,1)$
- SG (basic version): $\mathbb{E}[f(\boldsymbol{w}_k) - f(\boldsymbol{w}^\star)] \leq \mathcal{O}(1/k)$,

Heavy computation

- Large scale optimization, $N \gg 1$ for modern learning tasks

Heavy communication

- Large-scale learning solved with distributed optimization

Security

- Revealing only a noisy gradient information

Nonconvex optimization and saddle points

Generic SGD algorithm

Stochastic Gradient Descent (SGD) algorithm

Initialize $\boldsymbol{w}_1$
for $k = 1, 2, \ldots,$ **do**
  Generate a realization of the random variable ζ_k
  Compute a stochastic vector $g(\boldsymbol{w}_k; \zeta_k)$
  Choose step-size $\alpha_k > 0$
  Update $\boldsymbol{w}_{k+1} \leftarrow \boldsymbol{w}_k - \alpha_k \; g(\boldsymbol{w}_k; \zeta_k)$
end for

- Examples of stochastic vector, ζ_k

  Gradient for one sample, $\nabla f_{\zeta_k}(\boldsymbol{w}_k)$: Plain/Vanilla SGD

  Gradient for a mini-batch, $\frac{1}{N_k}\sum_{i\in[N_k]}\nabla f_{\zeta_k,i}(\boldsymbol{w}_k)$: Mini-batch SGD

  Preconditioned mini-batch gradient, $\boldsymbol{H}_k\left(\frac{1}{N_k}\sum_{i\in[N_k]}\nabla f_{\zeta_k,i}(\boldsymbol{w}_k)\right)$

Outline

Setup and Assumptions

$\boldsymbol{w}_{k+1}$: depends only on ζ_k, and **assume i.i.d.** $\{\zeta_k\}_k$
at iteration k, $\boldsymbol{w}_k$ deterministic.
randomness comes from $\boldsymbol{w}_{k+1}$: due to randomly selected batch ζ_k

$\mathbb{E}_{\zeta_k}[f(\boldsymbol{w}_{k+1})]$: expectation of $f(\boldsymbol{w}_{k+1})$ wrt the distribution of ζ_k only

Assume f to be L-**smooth**.

Assume $g(\boldsymbol{w}_k;\zeta_k)$ **unbiased estimator** of $\nabla f(\boldsymbol{w}_k)$: $\mathbb{E}_{\zeta_k}\left[g(\boldsymbol{w}_k;\zeta_k)\right] = \nabla f(\boldsymbol{w}_k)$

Bounded gradient: There exist scalars $c_0 \geq c > 0$ s.t. for all $k \in \mathbb{N}$

$$\frac{\|\mathbb{E}_{\zeta_k}\left[g(\boldsymbol{w}_k;\zeta_k)\right]\|_2^2}{c_0^2} \leq \|\nabla f(\boldsymbol{w}_k)\|_2^2 \leq \frac{\nabla f(\boldsymbol{w}_k)^T \mathbb{E}_{\zeta_k}\left[g(\boldsymbol{w}_k;\zeta_k)\right]}{c} \tag{1}$$

- RHS of (1): SG estimation within the half-space of true grad (on average), to have decrease
- LHS of (1): norm of error in SG estimate bounded by norm of true grad

For unbiased estimator: $c = 1$. why ?

Bounded variance: There exist scalars $M \geq 0$ and $M_V \geq 0$ s.t. for all $k \in \mathbb{N}$

$$\mathrm{Var}_{\zeta_k}\left[g(\boldsymbol{w}_k;\zeta_k)\right] \leq M + M_V \|\nabla f(\boldsymbol{w}_k)\|_2^2 \tag{2}$$

and $M_G = M_V + c_0^2$.

Analysis of single SGD step

$$
\begin{aligned}
&\mathbb{E}_{\zeta_k}[f(\boldsymbol{w}_{k+1})] - f(\boldsymbol{w}_k) \\
&\overset{(a)}{\leq} -\underbrace{\alpha_k \nabla f(\boldsymbol{w}_k)^T \mathbb{E}_{\zeta_k}\left[g(\boldsymbol{w}_k;\zeta_k)\right]}_{\text{expected decrease}} + \underbrace{\frac{1}{2}\alpha_k^2 L \mathbb{E}_{\zeta_k}\left[\|g(\boldsymbol{w}_k;\zeta_k)\|_2^2\right]}_{\text{noise}} \\
&\overset{(b)}{=} -\alpha_k \underbrace{\|\nabla f(\boldsymbol{w}_k)\|_2^2}_{\text{true grad norm}} + \frac{1}{2}\alpha_k^2 L \underbrace{\mathrm{Var}_{\zeta_k}\left[g(\boldsymbol{w}_k;\zeta_k)\right]}_{\text{var. of SG estimator}} \\
&\overset{(c)}{\leq} -\left(c - \frac{1}{2}\alpha_k L M_G\right) + \frac{1}{2}\alpha_k^2 L M
\end{aligned}
$$

- (b) follows from $g(\boldsymbol{w}_k;\zeta_k)$ is unbiased estiamtor of $\nabla\ f(\boldsymbol{w}_k)$
- (c) follows from bounds on gradients norm (1) and variance of SG (2)

Convergence of SG depends on the balance between blue and red terms:

- blue part helps to decrease the cost (good)
- variance of SG estimator (red) pushes in other direction
- **reducing variance (red) improves convergence**

Outline

Strongly convex f and fixed step-size

Theorem 1: Convergence of SGD for strongly convex f and constant step
For all $k \in \mathbb{N}$ and constant step-size $\alpha_k = \alpha$ satisfying

$$0 < \alpha \leq \frac{c}{LM_G}\,, \tag{3}$$

the expected optimality gap satisfies

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] &\leq \frac{\alpha LM}{2\mu c} + (1-\alpha\mu c)^{k-1}\left(f(\boldsymbol{w}_1) - f^\star - \frac{\alpha LM}{2\mu c}\right)\\ &\approx \frac{\alpha LM}{2\mu c} + \mathcal{O}(\rho^{k-1}) \ \xrightarrow{k\to\infty}\ \frac{\alpha LM}{2\mu c}\end{aligned} \tag{4}$$

▸ Proof

Strongly convex f and fixed step-size

$$\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] \leq (1-\alpha\mu c)^{k-1}\left(f(\boldsymbol{w}_1) - f^\star - \frac{\alpha L M}{2\mu c}\right) + \frac{\alpha L M}{2\mu c}$$

Fast convergence to a neighborhood of the optimal value, but noise in the gradient preventd further progress (convergence to an ambiguity ball)

Nonzero optimality gap $= \frac{\alpha L M}{2\mu c}$ independent of k

A simple modification: run SG with a fixed step-size, and after convergence half the step-size.

- How $E[f(\boldsymbol{w}_k)]$ against k behaves now?

- No sub-optimality gap

- Each time the step-size is cut in half, double the number of iterations are required

Strongly convex f and diminishing step-size

Theorem 2: Strongly convex f and diminishing stepsize
For all $k \in \mathbb{N}$ and diminishing step-size α_k satisfying

$$\alpha_k = \frac{\beta}{\gamma + k}, \text{ for some } \beta > \frac{1}{\mu c} \text{ and } \gamma > 0 \text{ s.t. } \alpha_1 \leq \frac{c}{L M_G},$$

the expected optimality gap satisfies

$$\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] \leq \frac{\nu}{\gamma + k} \approx \mathcal{O}(1/k) \tag{5}$$

where

$$\nu := \max\left\{\frac{\beta^2 L M}{2(\beta\mu c - 1)}, (\gamma + 1)\left(f(\boldsymbol{w}_1) - f^\star\right)\right\}$$

Optimality gap decays to zero. Similar convergence behavior as GD: $\mathcal{O}(1/k)$

need to known L, M_G

Proof

Strongly convex f and diminishing step-size

One may wish a **mix-and-match strategy**:

Constant step-size and mini-batch vs diminishing step-size

For mini-batch, define $g(\boldsymbol{w}_k;\zeta_k) = \frac{1}{N_m}\sum_{i\in[N_m]} \nabla f_{\zeta_k,i}(\boldsymbol{w}_k)$

SGD with mini-batch size of N_m and constant step

Mini-batch with small constant $\alpha > 0$,

$$\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] \le \frac{\alpha L M}{2\mu c N_m} + (1-\alpha\mu c)^{k-1}\left(f(\boldsymbol{w}_1) - f^\star - \frac{\alpha L M}{2\mu c N_m}\right)$$

optimality gap reduced by increasing mini-batch size

Compare with simple SG with "effective constant step-size" α/N_m,

$$\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] \le \frac{\alpha L M}{2\mu c N_m} + \left(1-\frac{\alpha\mu c}{N_m}\right)^{k-1}\left(f(\boldsymbol{w}_1) - f^\star - \frac{\alpha L M}{2\mu c N_m}\right)$$

slower convergence, but iterations are cheaper

More results on **convex f with diminishing stepsize** in supplements. Here

Non-convex objective function

Theorem 5: Convergence for non-convex function
With fixed step-size $0 < \alpha \leq c/(LM_G)$, for all $K \in \mathbb{N}$, we have

$$\mathbb{E}\left[\frac{1}{K}\sum_{k\in[K]}\|\nabla f(\boldsymbol{w}_k)\|_2^2\right] \leq \frac{\alpha LM}{c} + \frac{2(f(\boldsymbol{w}_1) - f_{\inf})}{Kc\alpha} \xrightarrow{K\to\infty} \frac{\alpha LM}{c} \tag{6}$$

Convergence criteria: average of squared gradient norm, over K samples

$f_{\inf} := \min_{k\in[K]}\ f(\boldsymbol{w}_k)$ (not necessarily $f^\star$)

SG spends increasingly more time in regions where the objective function has a "relatively" small gradient. Also usual tradeoff on step-size.

More results on **non-convex** f **with diminishing stepsize** in supplements.

Here

Outline

Intuition for Variance Reduction

Recall: Reducing variance of SG estimator improves convergence.
How to do that ?

Increasing mini-batch size

Reducing step-size α_k to reduce the radius of the ambiguity ball

Better trick: Given random variables, X, Y.
Assume that the expectation of Y, $\mathbb{E}[Y]$ known.

Reducing variance of X by defining $Z = a(X - Y) + \mathbb{E}[Y]$

It follows: $\mathbb{E}[Z] = a\mathbb{E}[X] + (1 - a)\mathbb{E}[Y]$

and $\mathrm{var}(Z) = a^2 \left(\mathrm{var}(X) + \mathrm{var}(Y) - 2\mathrm{cov}(X, Y)\right)$

$a = 1$: Z unbiased estimator of X

$a < 1$: $\mathrm{var}(Z) \leq \mathrm{var}(X)$ for some choices of Y

- reduce variance at the expense of increasing potential bias
- tradeoff b/w bias and variance in known in statistics and estimation

Stochastic variance reduced gradient (SVRG)

SVRG (Johnson&Zhang, 2013; Zhang et. al., 2013)

Inputs: Epoch length T, number of epochs K
Initialize $\widetilde{\boldsymbol{w}}_k$
for $k = 1, 2, \ldots, K$ **do**
  Compute full gradients and store $\widetilde{\nabla} f := \frac{1}{N} \sum_{i \in [N]} \nabla f_i(\widetilde{\boldsymbol{w}}_k)$
  Initialize $\boldsymbol{w}_{k,0} = \widetilde{\boldsymbol{w}}_k$
  for t=1,. . . ,T **do**
    Sample ζ_k uniformly from $[N]$
    $\boldsymbol{w}_{k,t} = \boldsymbol{w}_{k,t-1} - \alpha_k \left(\nabla f_{\zeta_k}(\boldsymbol{w}_{k,t-1}) - \nabla f_{\zeta_k}(\widetilde{\boldsymbol{w}}_k) + \widetilde{\nabla} f \right)$
  end for
  Update $\widetilde{\boldsymbol{w}}_{k+1} = \boldsymbol{w}_{k,T}$
end for
Return: $\widetilde{\boldsymbol{w}}_{K+1}$

- $X = \boldsymbol{w}$, $Z = \widetilde{\boldsymbol{w}}$, with a known average (blue step)
- Update $\widetilde{\boldsymbol{w}}_k$ in outer loop. Update $\boldsymbol{w}_{k,t}$ in inner loop
- cost two gradient computation per inner loop
- Linear convergence rate (given a sufficiently large T)

Proof

Stochastic average gradient (SAG)

SAG (Schmidt&Le Roux&Bach, 2012, 2017)

for $k = 1, 2, \ldots,$ **do**

Sample ζ_k uniformly from $[N]$ and observe $\nabla f_i(\boldsymbol{w}_k)$

Update for all $i \in [N]$, $\hat{g}_i(\boldsymbol{w}_k) = \begin{cases} \nabla f_i(\boldsymbol{w}_k), & \text{if } i = \zeta_k \\ \hat{g}_i(\boldsymbol{w}_{k-1}), & \text{otherwise} \end{cases}$

Update $\boldsymbol{w}_{k+1} \leftarrow \boldsymbol{w}_k - \frac{\alpha_k}{N} \sum_{i \in [N]} \hat{g}_i(\boldsymbol{w}_k)$

end for

- One main loop (like SGD)
- Almost same convergence rate (and same proof) as of SVRG
- A memory of size N
- Biased gradient estimates: $\mathbb{E}\left[\frac{1}{N} \sum_{i \in [N]} \hat{g}_i(\boldsymbol{w}_k)\right] = \frac{1}{N} \sum_{i \in [N]} \nabla f_i(\boldsymbol{w}_k)$ does not hold necessarily

Conclusions

Motivated and presented SGD as fundamental building block of large-scale learning

Presented convergence results for several functions (convex, strongly convex, and non-convex)

and different choices of stepsize: constant, decaying

Variance Reduction methods to reduce variance of SG estimator:

Stochastic variance reduced gradient (SVRG) and Stochastic Average Gradient (SAG)

Some references

- L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," SIAM Review, 2018.

- S. Bubeck, "Convex optimization: Algorithms and complexity," Foundations and Trends in Machine Learning, 2015.

- S. Boyd and A. Mutapcic, "Stochastic subgradient methods," Lecture Notes for EE364b, Stanford University, 2018.

- R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," NIPS, 2013.

- L. Zhang, M. Mahdavi, and R. Jin, "Linear convergence with condition number independent access of full gradients," NIPS 2013.

- M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," Mathematical Programming, 2017.

Outline

Convex f and diminishing step-size

- Notations:
  - $\mathbb{E}[g(\boldsymbol{w};\zeta_k)|\boldsymbol{w}_k] \in \partial f(\boldsymbol{w}_k)$: noisy unbiased sub-gradient of convex f
  - $f_{\text{best}}(\boldsymbol{w}_k) = \min\left(f(\boldsymbol{w}_1), \ldots, f(\boldsymbol{w}_k)\right)$
  - $\mathbb{E}\left[\|g(\boldsymbol{w}_k;\zeta_k)\|_2^2\right] \leq G^2$ for all k, and $\sup_{\boldsymbol{w}\in\mathcal{W}} \mathbb{E}\left[\|\boldsymbol{w}_1 - \boldsymbol{w}^\star\|_2^2\right] \leq R^2$

Theorem 3

Under some mild conditions and for square summable but not summable step-size, we have convergence in expectation

$$\mathbb{E}\left[f_{\text{best}}(\boldsymbol{w}_k) - f^\star\right] \leq \frac{R^2 + G^2 \sum_{i\in[k]} \alpha_i^2}{2\sum_{i\in[k]} \alpha_i}$$

and for any arbitrary $\epsilon, \delta > 0$, we have convergence in probability:

$$\Pr\left(f_{\text{best}}(\boldsymbol{w}_k) - f^\star \geq \epsilon\right) \leq \delta$$

Proof

Convex f and diminishing step-size

Theorem 4

For convex L-smooth function f, i.i.d. stochastic gradient of variance bound σ^2, and diminishing step-size $\alpha_k = \frac{1}{L+\gamma^{-1}}$, where $\gamma = \frac{R}{G}\sqrt{\frac{2}{k}}$, we have

$$\mathbb{E}\left[f\left(\frac{1}{k}\sum_{i\in[k]} \boldsymbol{w}_k\right) - f^\star\right] \leq R\sqrt{\frac{2\sigma^2}{k}} + \frac{LR^2}{k} \tag{7}$$

Proof: see [Bubeck 2015, Theorem 6.3]

Improved gain for mini-batch of size N_m: $\sigma^2 \to \sigma^2/N_m$

Proof sketch for Theorem 1

Use (**??**), Polyak-Lojasiewicz inequality (a consequence of strong convexity) and (3), and observe that

$$\begin{aligned}\mathbb{E}_{\zeta_k}[f(\boldsymbol{w}_{k+1})] - f(\boldsymbol{w}_k) &\leq -\left(c - \frac{1}{2}\alpha L M_G\right)\alpha\|\nabla f(\boldsymbol{w}_k)\|_2^2 + \frac{1}{2}\alpha^2 L M \\ &\leq -\frac{1}{2}\alpha c\|\nabla f(\boldsymbol{w}_k)\|_2^2 + \frac{1}{2}\alpha^2 L M \\ &\leq -\alpha\mu c\,(f(\boldsymbol{w}_k) - f^\star) + \frac{1}{2}\alpha^2 L M\end{aligned}$$

Subtract $f^\star$ from both sides, take total expectation, and rearrange:

$$\mathbb{E}\left[f(\boldsymbol{w}_{k+1}) - f^\star\right] \leq (1 - \alpha\mu c)\,\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] + \frac{1}{2}\alpha^2 L M$$

Make it a contraction inequality (as $0 < \alpha\mu c \leq \frac{\mu c^2}{L M_G} \leq \frac{\mu}{L} \leq 1$)

$$\mathbb{E}\left[f(\boldsymbol{w}_{k+1}) - f^\star\right] - \frac{\alpha L M}{2\mu c} \leq (1 - \alpha\mu c)\left(\mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] - \frac{\alpha L M}{2\mu c}\right).$$

Return

Non-convex objective function

Theorem 5

With square summable but not summable step-size, we have for any $K \in \mathbb{N}$

$$\mathbb{E}\left[\sum_{k\in[K]} \alpha_k \|\nabla f(\boldsymbol{w}_k)\|_2^2\right] < \infty \tag{8}$$

and therefore

$$\mathbb{E}\left[\frac{1}{\sum_{k\in[K]} a_k} \sum_{k\in[K]} \alpha_k \|\nabla f(\boldsymbol{w}_k)\|_2^2\right] \xrightarrow{K\to\infty} 0 \tag{9}$$

The expected gradient norm cannot stay bounded away from zero

Proof sketch for Theorem 2

First observe that $\alpha_k L M_G \leq \alpha_1 L M_G \leq c$. Use (**??**) and Polyak-Lojasiewicz inequality and show that

$$\mathbb{E}_{\zeta_k}[f(\boldsymbol{w}_{k+1})] - f(\boldsymbol{w}_k) \leq -\alpha_k \mu c \left(f(\boldsymbol{w}_k) - f^\star\right) + \frac{1}{2}\alpha_k^2 L M$$

Subtract $f^\star$ from both sides, take total expectation, and rearrange:

$$\mathbb{E}\left[f(\boldsymbol{w}_{k+1}) - f^\star\right] \leq (1 - \alpha_k \mu c)\, \mathbb{E}\left[f(\boldsymbol{w}_k) - f^\star\right] + \frac{1}{2}\alpha_k^2 L M$$

Now prove by induction and use inequality $k^2 \geq (k+1)(k-1)$

▸ Return

Proof sketch for Theorem 3

Use convexity of f $\left(f^\star - f(\boldsymbol{w}_k) \geq \mathbb{E}[g(\boldsymbol{w};\zeta_k)|\boldsymbol{w}_k]^T(\boldsymbol{w}^\star - \boldsymbol{w}_k)\right)$ to show

$$\mathbb{E}\left[\|\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star\|_2^2 \mid \boldsymbol{w}_k\right] \leq \|\boldsymbol{w}_k - \boldsymbol{w}^\star\|_2^2 - 2\alpha_k\left(f(\boldsymbol{w}_k) - f^\star\right) + \alpha_k^2 G^2$$

Take expectation nd apply recursively to show

$$\mathbb{E}\left[\|\boldsymbol{w}_{k+1} - \boldsymbol{w}^\star\|_2^2\right] \leq \mathbb{E}[\|\boldsymbol{w}_1 - \boldsymbol{w}^\star\|_2^2] - 2\sum_{i\in[k]}\alpha_i\left(\mathbb{E}[f(\boldsymbol{w}_i)] - f^\star\right) + G^2\sum_{i\in[k]}\alpha_i^2$$

Conclude that for square summable but not summable step-size, $\min_{i\in[k]}\mathbb{E}[f(\boldsymbol{w}_i)] \to f^\star$

Use Jensen's inequality and concavity of minimum to show convergence in expectation $\mathbb{E}[f_{\text{best}}(\boldsymbol{w}_k)] = \mathbb{E}[\min_{i\in[k]} f(\boldsymbol{w}_i)] \leq \min_{i\in[k]}\mathbb{E}[f(\boldsymbol{w}_i)] \to f^\star$

Use Markov's inequality to show convergence in probability: $\Pr(f_{\text{best}}(\boldsymbol{w}_k) - f^\star \geq \epsilon) \leq \frac{\mathbb{E}[f_{\text{best}}(\boldsymbol{w}_k) - f^\star]}{\epsilon}$

Return

Linear convergence of SVRG

Variance decomposition:

$$\mathbb{E}\left[\|\boldsymbol{w} - \mathbb{E}[\boldsymbol{w}]\|_2^2\right] \le \mathbb{E}\left[\|\boldsymbol{w}\|_2^2\right] - \|\mathbb{E}[\boldsymbol{w}]\|_2^2 \le \mathbb{E}\left[\|\boldsymbol{w}\|_2^2\right]$$

Show

$$\mathbb{E}_{\zeta_k}\left[\left\|\nabla f_{\zeta_k}(\boldsymbol{w}_{k,t-1}) - \nabla f_{\zeta_k}(\widetilde{\boldsymbol{w}}_k) + \widetilde{\nabla} f\right\|_2^2\right] \le 4L\left(f(\boldsymbol{w}_{k,t-1}) + f(\widetilde{\boldsymbol{w}}_k) - 2f^\star\right)$$

Use the inner-loop iteration and bound $\mathbb{E}_{\zeta_k}\left[\|\boldsymbol{w}_{k,t} - \boldsymbol{w}^\star\|_2^2\right]$. You may need to use convexity of f

Sum $\mathbb{E}_{\zeta_k}\left[\|\boldsymbol{w}_{k,t} - \boldsymbol{w}^\star\|_2^2\right]$ over the inner loop ($t \in [T]$) and cancel some terms from both sides

Show and use for every outer iteration to observe the linear convergence rate: if $a < ba + c$ for $b \in (0,1)$, then

$$a - \frac{c}{1-b} \le b\left(a - \frac{c}{1-b}\right)$$

Return