CMPT 409/981: Optimization for Machine Learning

Lecture 12

Sharan Vaswani

October 27, 2022

Recap

When minimizing smooth, strongly-convex functions under interpolation, SGD with $\eta = \frac{1}{L}$ can converge to the minimizer at an $O(\exp(-{}^{T}\!/\!_{\kappa}))$ rate. Hence, SGD matches the rate of deterministic GD, but compared to GD, each iteration is cheap.

For smooth, non-convex functions, we require a stronger condition: $\mathbb{E}_i \left\|\nabla f_i(w)\right\|^2 \leq \rho \left\|\nabla f(w)\right\|^2$ under which constant step-size SGD can attain the deterministic GD rate.

To simplify the proofs, we considered that $f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$, but neither the algorithm nor the theoretical results depend on this finite-sum structure. We can extend all SGD results to handle a general stochastic oracle that returns $\nabla f(w, \zeta)$ s.t. $\mathbb{E}_\zeta[\nabla f(w, \zeta)] = \nabla f(w)$.

Stochastic Line-Search

Algorithmically, convergence under interpolation requires knowledge of L. We will use a *stochastic line-search* (SLS) procedure [VML$^+$19] to estimate L. SLS is similar to the deterministic variant in Lecture 3, but uses only stochastic function/gradient evaluations.

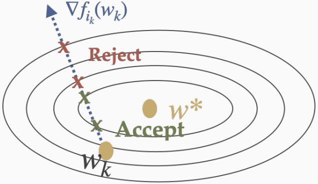
Algorithm SGD with Stochastic Line-search

```
function SGD with Stochastic Line-search ($f$, $w_0$, $\eta_{\max}$, $c \in (0, 1)$, $\beta \in (0, 1)$)
for $k = 0, \ldots, T - 1$ do
    $\tilde{\eta}_k \leftarrow \eta_{\max}$
    while $f_{ik}(w_k - \tilde{\eta}_k \nabla f_{ik}(w_k)) > f_{ik}(w_k) - c \cdot \tilde{\eta}_k \left\|\nabla f_{ik}(w_k)\right\|^2$ do
        $\tilde{\eta}_k \leftarrow \tilde{\eta}_k \beta$
    end while
    $\eta_k \leftarrow \tilde{\eta}_k$
    $w_{k+1} = w_k - \eta_k \nabla f_{ik}(w_k)$
end for
return $w_T$
```

Stochastic Line-Search

SLS searches for a good step-size in the wrong direction.

Since all f_i have zero gradient at w^* and the noise decreases as we get closer to the solution (because of interpolation), SGD with SLS converges to the minimizer.



Claim: If each f_i is L-smooth, then the (exact) backtracking procedure for SLS terminates and returns $\eta_k \in \left[\min\left\{\frac{2(1-c)}{L}, \eta_{\max}\right\}, \eta_{\max}\right]$.

Proof: Similar to the deterministic case (Lecture 3), but requires that each f_i is L-smooth.

Minimizing smooth, strongly-convex functions using SGD + SLS under interpolation

Claim: When minimizing $f(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$ such that (i) f is μ-strongly convex, (ii) each f_i is convex and L-smooth, (iii) interpolation is exactly satisfied i.e. $\|\nabla f_i(w^*)\| = 0$, T iterations of SGD with SLS (with $c = 1/2$) returns iterate w_T such that,

$$\mathbb{E}[\|w_T - w^*\|^2] \leq \exp\left(-\mu \, T \, \min\left\{\frac{1}{L}, \eta_{\max}\right\}\right) \|w_0 - w^*\|^2$$

Proof: Similar to the previous proof, we get that,

$$\mathbb{E}[\|w_{k+1} - w^*\|^2] = \|w_k - w^*\|^2 - 2\mathbb{E}\left[\eta_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle\right] + \mathbb{E}\left[\eta_k^2 \, \|\nabla f_{ik}(w_k)\|^2\right] \quad (1)$$

Since η_k depends on i_k, we can not push the expectation in. η_k is set by SLS, it satisfies the stochastic Armijo condition. Simplifying the third term and denoting $f_{ik}^* := \min f_{ik}(w)$,

$$\mathbb{E}\left[\eta_k^2 \, \|\nabla f_{ik}(w_k)\|^2\right] \leq \mathbb{E}\left[\eta_k \, \frac{f_{ik}(w_k) - f_{ik}(w_{k+1})}{c}\right] \leq \mathbb{E}\left[\eta_k \, \frac{f_{ik}(w_k) - f_{ik}^*}{c}\right] \quad (2)$$

Minimizing smooth, strongly-convex functions using SGD + SLS under interpolation

Using Eq. (1) + Eq. (2),

$$\mathbb{E}[\|w_{k+1} - w^*\|^2] = \|w_k - w^*\|^2 - 2\mathbb{E}\left[\eta_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle\right] + \mathbb{E}\left[\eta_k \frac{f_{ik}(w_k) - f_{ik}^*}{c}\right] \tag{3}$$

$$\begin{aligned}
\mathbb{E}\left[\eta_k \frac{f_{ik}(w_k) - f_{ik}^*}{c}\right] &= \mathbb{E}\left[2\eta_k \left(f_{ik}(w_k) - f_{ik}(w^*) + f_{ik}(w^*) - f_{ik}^*\right)\right] && \text{(Setting } c = 1/2\text{)} \\
&= \mathbb{E}\left[2\eta_k \left(f_{ik}(w_k) - f_{ik}(w^*)\right)\right] + \mathbb{E}\left[2\eta_k \underbrace{\left(f_{ik}(w^*) - f_{ik}^*\right)}_{\text{Positive}}\right] \\
&\leq \mathbb{E}\left[2\eta_k \left(f_{ik}(w_k) - f_{ik}(w^*)\right)\right] + 2\eta_{\max}\,\mathbb{E}\left[f_{ik}(w^*) - f_{ik}^*\right] && \text{(Since } \eta_k \leq \eta_{\max}\text{)}
\end{aligned}$$

Since f_{ik} is convex and $\nabla f_{ik}(w^*) = 0$, $f_{ik}(w^*) = f_{ik}^*$.

$$\mathbb{E}\left[\eta_k \frac{f_{ik}(w_k) - f_{ik}^*}{c}\right] \leq \mathbb{E}\left[2\eta_k \left(f_{ik}(w_k) - f_{ik}(w^*)\right)\right] \tag{4}$$

Minimizing smooth, strongly-convex functions using SGD + SLS under interpolation

Using Eq. (3) + Eq. (4),

$$
\begin{aligned}
\mathbb{E}[\|w_{k+1} - w^*\|^2] &= \|w_k - w^*\|^2 - 2\mathbb{E}\left[\eta_k \langle \nabla f_{ik}(w_k), w_k - w^* \rangle\right] + \mathbb{E}\left[2\eta_k \left(f_{ik}(w_k) - f_{ik}(w^*)\right)\right] \\
&= \|w_k - w^*\|^2 + 2\mathbb{E}\left[\eta_k (f_{ik}(w_k) - f_{ik}(w^*) + \langle \nabla f_{ik}(w_k), w^* - w_k \rangle)\right]
\end{aligned}
$$

Since f_{ik} is convex, $f_{ik}(w_k) - f_{ik}(w^*) + \langle \nabla f_{ik}(w_k), w^* - w_k \rangle \leq 0$

$$
\begin{aligned}
&\leq \|w_k - w^*\|^2 + 2\eta_{\min}\, \mathbb{E}\left[f_{ik}(w_k) - f_{ik}(w^*) + \langle \nabla f_{ik}(w_k), w^* - w_k \rangle\right] \\
&\qquad \text{(Lower-bounding } \eta_k \text{. } \eta_{\min} := \min\left\{\tfrac{1}{L}, \eta_{\max}\right\}\text{)} \\
&= \|w_k - w^*\|^2 + 2\eta_{\min}\, \mathbb{E}\left[f(w_k) - f(w^*) + \langle \nabla f(w_k), w^* - w_k \rangle\right] \\
&\qquad \text{(Unbiasedness)} \\
&\leq \|w_k - w^*\|^2 + 2\eta_{\min} \left[\frac{-\mu}{2} \|w_k - w^*\|^2\right] \qquad (f \text{ is } \mu\text{-strongly convex})
\end{aligned}
$$

$$
\implies \mathbb{E}[\|w_{k+1} - w^*\|^2] \leq (1 - \mu\, \eta_{\min})\, \|w_k - w^*\|^2
$$

Minimizing smooth, strongly-convex functions using SGD + SLS under interpolation

Recall that $\mathbb{E}[\|w_{k+1} - w^*\|^2] \leq (1 - \mu\, \eta_{\min})\ \|w_k - w^*\|^2$. Taking expectation w.r.t the randomness from iterations $k = 0$ to $T - 1$ and recursing,

$$\mathbb{E}[\|w_T - w^*\|^2] \leq (1 - \mu\eta_{\min})^T\ \|w_0 - w^*\|^2 \leq \exp\left(-\mu\, T\, \eta_{\min}\right)\ \|w_0 - w^*\|^2$$

$$\implies \mathbb{E}[\|w_T - w^*\|^2] \leq \exp\left(-\mu\, T\, \min\left\{\frac{1}{L}, \eta_{\max}\right\}\right)\ \|w_0 - w^*\|^2$$

Hence, when minimizing smooth, strongly-convex functions under interpolation, SGD + SLS will will converge to the minimizer at an exponential rate.

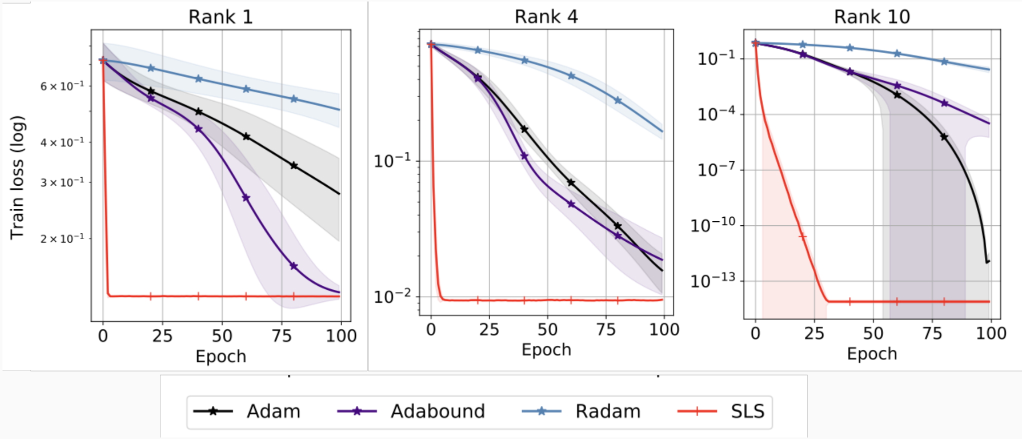
If interpolation is not exactly satisfied, we can modify the proof to get an $O\left(\exp\left(\frac{-T}{\kappa}\right) + \zeta^2\right)$ rate, where $\zeta^2 := \mathbb{E}\left[f_{ik}(w^*) - f_{ik}^*\right]$.

When minimizing convex functions under (exact) interpolation, SGD + SLS results in an $O(1/T)$ rate without requiring knowledge of L. (Need to prove this in Assignment 3!)

Do not have strong theoretical results for SGD + SLS on smooth, non-convex problems.

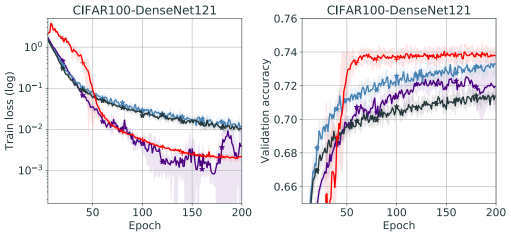
Stochastic Line-Search and Effect of Over-parametrization

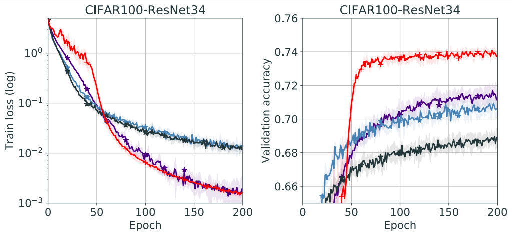
Objective: $\min_{\theta_1,\theta_2} \frac{1}{2n} \sum_{i=1}^{n} \|\theta_2\, \theta_1 x_i - y_i\|^2$; **Parameterization**: $\theta_1 \in \mathbb{R}^{k\times 6}$, $\theta_2 \in \mathbb{R}^{10\times k}$.



Stochastic Line-Search - Experimental Results

Task: Multi-class classification with logistic loss.





Stochastic Polyak Step-size

When interpolation is (approximately) satisfied, we can use SGD with the *stochastic Polyak step-size* (SPS) [LVLLJ21]: At iteration k, for hyper-parameter $c \in (0, 1)$ and $f_{ik}^* := \min_w f_{ik}(w)$,

$$\eta_k = \frac{f_{ik}(w_k) - f_{ik}^*}{c \left\|\nabla f_{ik}(w_k)\right\|^2} .$$

Common machine learning losses (squared loss, logistic loss, exponential loss) are lower-bounded by zero. Algorithmically, we can set $f_{ik}^* = 0$.

- SPS matches the SLS rates on smooth, (strongly) convex functions. E.g. SPS with $c = 1/2$ achieves the $O\left(\exp\left(\frac{-T}{\kappa}\right) + \zeta^2\right)$ rate for smooth, strongly-convex functions.
- Much simpler and computationally inexpensive to implement compared to SLS.
- Unlike SLS, SPS can be used for minimizing non-smooth, convex functions.
- Results in large step-sizes and requires some additional heuristics for stabilizing the method.
- For neural networks, generalization for SGD + SPS was typically worse than for SGD + SLS.
- Requires access to f_{ik}^* which might be difficult to compute for more general problems.

Adaptivity for SGD

Noise-adaptivity: When minimizing smooth, strongly-convex functions, with T iterations of SGD with $\eta_k := \frac{1}{L}\left(\frac{1}{T}\right)^{\frac{k}{T}}$, we can obtain an $O\left(\exp\left(\frac{-T}{\kappa}\right) + \frac{\zeta^2}{T}\right)$ rate, where $\zeta^2 := \mathbb{E}_i[f_i(w^*) - f_i^*]$. Adaptive to the extent of interpolation, but requires L to set the step-size.

Problem-adaptivity: SGD with the step-size set according to SLS/SPS is adaptive to L, but results in an $O\left(\exp\left(\frac{-T}{\kappa}\right) + \zeta^2\right)$ rate.

[VDTB21] attempts to combine the above ideas to obtain both noise and problem adaptivity i.e. use SLS to set $\gamma_k \approx \frac{1}{L}$ and use $\eta_k = \gamma_k \left(\frac{1}{T}\right)^{\frac{k}{T}}$. Either not guaranteed to converge to the minimizer or will converge to the minimizer at a slower (than optimal) rate.

For smooth, strongly-convex problems, we do not (yet) know how to make SGD problem and noise-adaptive, and achieve the optimal rate.

For smooth, convex problems, AdaGrad is both problem and noise-adaptive.

Questions?

Minimizing smooth, strongly-convex functions

For minimizing smooth, strongly-convex functions $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ to an ϵ-suboptimality,

- Deterministic GD requires $O(\kappa \log(1/\epsilon))$ iterations, and $O(n\,\kappa \log(1/\epsilon))$ gradient evaluations.
- SGD with a decreasing step-size requires $O(1/\epsilon)$ iterations, and $O(1/\epsilon)$ gradient evaluations.
- Under exact interpolation, SGD with a constant step-size requires $O(\kappa \log(1/\epsilon))$ iterations, and $O(\kappa \log(1/\epsilon))$ gradient evaluations.
- For finite-sum problems of the form $\frac{1}{n}\sum_{i=1}^{n} f_i(w)$, **variance reduced methods** require $O((n+\kappa)\log(1/\epsilon))$ gradient evaluations.

Variance Reduced Methods

Recall that under exact interpolation, the variance decreases as we approach the minimizer.

On the other hand, variance reduced methods explicitly reduce the variance by either storing the past stochastic gradients to approximate the full gradient [SLRB17] or by computing the full gradient every “few” iterations [JZ13].

With variance reduction, we can use acceleration techniques to improve the dependence on the condition number, and require $O((n + \sqrt{\kappa}) \log(1/\epsilon))$ gradient evaluations [AZ17].

For smooth, convex finite-sum problems, variance reduced techniques require $O\left((n + \frac{1}{\epsilon}) \log(1/\epsilon)\right)$ gradient evaluations [NLST17], compared to deterministic GD that requires $O(\frac{n}{\epsilon})$ gradient evaluations and SGD that requires $O(\frac{1}{\epsilon^2})$ gradient evaluations.

We will use SVRG (Stochastic Variance Reduced Gradient) [JZ13] for smooth, strongly-convex finite-sum problems, and prove that it requires $O((n + \kappa) \log(1/\epsilon))$ gradient evaluations.

SVRG

For simplicity, we will use Loopless SVRG [KHR20] that has a simpler implementation and analysis compared to the original paper [JZ13].

Algorithm SVRG

```
function SVRG ($f$, $w_0$, $\eta$, $p \in (0,1]$)
$v_0 = w_0$
for $k = 0, \ldots, T-1$ do
    $g_k = \nabla f_{ik}(w_k) - \nabla f_{ik}(v_k) + \nabla f(v_k)$
    $w_{k+1} = w_k - \eta g_k$
    $v_{k+1} = \begin{cases} v_k \text{ with probability } 1-p \\ w_k \text{ with probability } p \end{cases}$
end for
return $w_T$
```

Minimizing smooth, strongly-convex functions using SVRG

Claim: When minimizing $f(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$ such that (i) f is μ-strongly convex, (ii) each f_i is convex and L-smooth, T iterations of SVRG with $\eta = \frac{1}{6L}$ and $p = \frac{1}{n}$ returns iterate w_T,

$$\mathbb{E}[\|w_T - w^*\|^2] \leq \left(\max\left\{\left(1 - \frac{\mu}{6L}\right), \left(1 - \frac{1}{2n}\right)\right\}\right)^T \left[2n\,\|w_0 - w^*\|^2\right].$$

Case 1: $(1 - \frac{\mu}{6L}) \leq (1 - \frac{1}{2n}) \implies n \geq 3\kappa$. In this case, for achieving an ϵ-suboptimality, we need T iterations such that $T \geq 2n \log\left(\frac{2n\,\|w_0 - w^*\|^2}{\epsilon}\right)$.

Case 2: $(1 - \frac{\mu}{6L}) > (1 - \frac{1}{2n}) \implies n \leq 3\kappa$. In this case, for achieving an ϵ-suboptimality, we need T iterations such that $T \geq 6\kappa \log\left(\frac{2n\,\|w_0 - w^*\|^2}{\epsilon}\right)$.

Putting the cases together, for achieving an ϵ-suboptimality, we need $T = O\left((n + \kappa) \log(1/\epsilon)\right)$.

In each iteration, the number of expected gradient evaluations is $(1 - p)(2) + (p)(n + 2) = pn + 2 = 3$. Hence, in expectation, SVRG requires $O\left((n + \kappa) \log(1/\epsilon)\right)$ gradient evaluations to achieve an ϵ-suboptimality.

Minimizing smooth, strongly-convex functions using SVRG

Proof: Using the algorithm update, $w_{k+1} = w_k - \eta g_k$ and following a similar proof as before,

$$
\begin{aligned}
\|w_{k+1} - w^*\|^2 &= \|w_k - w^*\|^2 - 2\eta \langle g_k, w_k - w^* \rangle + \eta^2 \|g_k\|^2 \\
\implies \mathbb{E}\|w_{k+1} - w^*\|^2 &= \|w_k - w^*\|^2 - 2\eta \langle \mathbb{E}[g_k], w_k - w^* \rangle + \eta^2\, \mathbb{E}[\|g_k\|^2] \\
&\qquad\qquad \text{(Since } \eta \text{ does not depend on } i_k\text{)} \\
&= \|w_k - w^*\|^2 - 2\eta \langle \nabla f(w_k), w_k - w^* \rangle + \eta^2\, \mathbb{E}[\|g_k\|^2] \\
&\qquad (\mathbb{E}[g_k] = \mathbb{E}[\nabla f_{ik}(w_k) - \nabla f_{ik}(v_k) + \nabla f(v_k)] = \nabla f(w_k))
\end{aligned}
$$

By strong-convexity,

$$
\mathbb{E}\|w_{k+1} - w^*\|^2 \leq (1 - \mu\eta)\|w_k - w^*\|^2 - 2\eta\,[f(w_k) - f(w^*)] + \eta^2\, \mathbb{E}[\|g_k\|^2] \tag{5}
$$

Next, we will bound $\mathbb{E}[\|g_k\|^2]$.

Minimizing smooth, strongly-convex functions using SVRG

$$
\begin{aligned}
\mathbb{E}[\|g_k\|^2] &= \mathbb{E}[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(v_k) + \nabla f(v_k)\|^2] \\
&= \mathbb{E}[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*) + \nabla f_{ik}(w^*) - \nabla f_{ik}(v_k) + \nabla f(v_k)\|^2] \\
&\leq 2\mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*)\|^2\right] + 2\mathbb{E}\left[\|\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k) + \nabla f(v_k)\|^2\right] \\
&\qquad\qquad ((a+b)^2 \leq 2a^2 + 2b^2) \\
&= 2\mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*)\|^2\right] + 2\mathbb{E}\left[\|\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k) - \mathbb{E}\left[\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k)\right]\|^2\right] \\
&\qquad\qquad (\text{Since } \mathbb{E}[\nabla f_{ik}(w^*)] = \nabla f(w^*) = 0)
\end{aligned}
$$

For any vector x, $\mathbb{E}\left[\|x - \mathbb{E}[x]\|^2\right] \leq \mathbb{E}[\|x\|^2]$. Using this with $x = \nabla f_{ik}(w^*) - \nabla f_{ik}(v_k)$

$$
\begin{aligned}
&\leq 2\mathbb{E}\left[\|\nabla f_{ik}(w_k) - \nabla f_{ik}(w^*)\|^2\right] + 2\mathbb{E}\left[\|\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k)\|^2\right] \\
&\leq 4L\,\mathbb{E}\left[f_{ik}(w_k) - f_{ik}(w^*) + \langle \nabla f_{ik}(w^*), w^* - w_k \rangle\right] + 2\mathbb{E}\left[\|\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k)\|^2\right] \\
&\qquad\qquad (\text{Smoothness of } f_{ik})
\end{aligned}
$$

$$
\implies \mathbb{E}[\|g_k\|^2] \leq 4L\,\mathbb{E}[f(w_k) - f(w^*)] + 2\mathbb{E}\left[\|\nabla f_{ik}(w^*) - \nabla f_{ik}(v_k)\|^2\right] \qquad (6)
$$

Minimizing smooth, strongly-convex functions using SVRG

Using Eq. (5) with Eq. (6),

$$\begin{aligned}\mathbb{E}\left\|w_{k+1}-w^*\right\|^2 &\leq (1-\mu\eta)\left\|w_k-w^*\right\|^2 - 2\eta\left[f(w_k)-f(w^*)\right] \\ &\quad + \eta^2\left[4L\,\mathbb{E}[f(w_k)-f(w^*)] + 2\mathbb{E}\left[\left\|\nabla f_{ik}(w^*)-\nabla f_{ik}(v_k)\right\|^2\right]\right] \\ &\leq (1-\mu\eta)\left\|w_k-w^*\right\|^2 + (4L\eta^2-2\eta)\,\mathbb{E}\left[f(w_k)-f(w^*)\right] \\ &\quad + \frac{2\eta^2}{n}\sum_{i=1}^{n}\left[\left\|\nabla f_i(w^*)-\nabla f_i(v_k)\right\|^2\right]\end{aligned}$$

Define $\mathcal{D}_k := \frac{4\eta^2}{pn}\sum_{i=1}^{n}\left[\left\|\nabla f_i(w^*)-\nabla f_i(v_k)\right\|^2\right]$.

$$\mathbb{E}\left\|w_{k+1}-w^*\right\|^2 \leq (1-\mu\eta)\left\|w_k-w^*\right\|^2 + (4L\eta^2-2\eta)\,\mathbb{E}\left[f(w_k)-f(w^*)\right] + \frac{p}{2}\mathcal{D}_k \qquad (7)$$

Minimizing smooth, strongly-convex functions using SVRG

Recall that $\mathcal{D}_k = \frac{4\eta^2}{pn} \sum_{i=1}^{n} \left[\|\nabla f_i(w^*) - \nabla f_i(v_k)\|^2\right]$. Using the algorithm,

$$\begin{aligned}
\mathbb{E}[\mathcal{D}_{k+1}] &= (1-p)\mathcal{D}_k + p\,\frac{4\eta^2}{pn}\sum_{i=1}^{n}\left[\|\nabla f_i(w^*) - \nabla f_i(w_k)\|^2\right] \\
&\leq (1-p)\mathcal{D}_k + \frac{8\eta^2\,L}{n}\sum_{i=1}^{n}\left[f_i(w_k) - f_i(w^*) + \langle \nabla f_i(w^*), w^* - w_k\rangle\right] && \text{(Smoothness)}
\end{aligned}$$

$$\implies \mathbb{E}[\mathcal{D}_{k+1}] \leq (1-p)\mathcal{D}_k + 8\eta^2\,L\,[f(w_k) - f(w^*)] \tag{8}$$

Minimizing smooth, strongly-convex functions using SVRG

Using Eq. (7) + Eq. (8),

$$\begin{aligned}
\mathbb{E}\left\|w_{k+1}-w^*\right\|^2+\mathbb{E}[\mathcal{D}_{k+1}] &\leq (1-\mu\eta)\left\|w_k-w^*\right\|^2+(4L\eta^2-2\eta)\,\mathbb{E}\left[f(w_k)-f(w^*)\right]+\frac{p}{2}\mathcal{D}_k \\
&\quad +(1-p)\mathcal{D}_k+8\eta^2 L\left[f(w_k)-f(w^*)\right] \\
&= (1-\mu\eta)\left\|w_k-w^*\right\|^2+(12L\eta^2-2\eta)\left[f(w_k)-f(w^*)\right]+\left(1-\frac{p}{2}\right)\mathcal{D}_k \\
&= \left(1-\frac{\mu}{6L}\right)\left\|w_k-w^*\right\|^2+\left(1-\frac{p}{2}\right)\mathcal{D}_k \qquad \text{(Since } \eta=\tfrac{1}{6L}\text{)} \\
&\leq \max\left\{\left(1-\frac{\mu}{6L}\right),\left(1-\frac{p}{2}\right)\right\}\left[\left\|w_k-w^*\right\|^2+\mathcal{D}_k\right] \\
\mathbb{E}\left[\left\|w_{k+1}-w^*\right\|^2+\mathcal{D}_{k+1}\right] &\leq \max\left\{\left(1-\frac{\mu}{6L}\right),\left(1-\frac{1}{2n}\right)\right\}\left[\left\|w_k-w^*\right\|^2+\mathcal{D}_k\right] \\
&\qquad \text{(Since } p=\tfrac{1}{n}\text{)}
\end{aligned}$$

Define $\Phi_k := \left[\left\|w_k-w^*\right\|^2+\mathcal{D}_k\right]$ and $\rho := \max\left\{\left(1-\frac{\mu}{6L}\right),\left(1-\frac{1}{2n}\right)\right\}$

$$\implies \mathbb{E}[\Phi_{k+1}] \leq \rho\,\Phi_k$$

Minimizing smooth, strongly-convex functions using SVRG

Recall that $\mathbb{E}[\Phi_{k+1}] \le \rho \Phi_k$. Taking expectation w.r.t the randomness in iterations from $k = 0$ to $T - 1$ and recursing,

$$\mathbb{E}[\Phi_T] \le \rho^T \Phi_0$$

$$\implies \mathbb{E}[\|w_T - w^*\|^2] \le \rho^T \left[\|w_0 - w^*\|^2 + \mathcal{D}_0\right] \quad \text{(Lower bounding } \phi_T \text{ since } \mathcal{D}_T \text{ is positive)}$$

$$= \rho^T \left[\|w_0 - w^*\|^2 + 4\eta^2 \sum_{i=1}^{n} \|\nabla f_i(w_0) - \nabla f_i(w^*)\|^2\right]$$

$$\le \rho^T \left[\|w_0 - w^*\|^2 + 4\eta^2 L^2 \sum_{i=1}^{n} \|w_0 - w^*\|^2\right] \quad \text{(Smoothness)}$$

$$\implies \mathbb{E}[\|w_T - w^*\|^2] \le \left(\max\left\{\left(1 - \frac{\mu}{6L}\right), \left(1 - \frac{1}{2n}\right)\right\}\right)^T \left[2n\|w_0 - w^*\|^2\right]$$

$$\text{(Since } \eta = \tfrac{1}{6L}\text{)}$$

Questions?

References i

- Zeyuan Allen-Zhu, *Katyusha: The first direct acceleration of stochastic gradient methods*, The Journal of Machine Learning Research **18** (2017), no. 1, 8194–8244.
- Rie Johnson and Tong Zhang, *Accelerating stochastic gradient descent using predictive variance reduction*, Advances in neural information processing systems **26** (2013).
- Dmitry Kovalev, Samuel Horváth, and Peter Richtárik, *Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop*, Algorithmic Learning Theory, PMLR, 2020, pp. 451–467.
- Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien, *Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence*, International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 1306–1314.

References ii

- Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč, *Sarah: A novel method for machine learning problems using stochastic recursive gradient*, International Conference on Machine Learning, PMLR, 2017, pp. 2613–2621.
- Mark Schmidt, Nicolas Le Roux, and Francis Bach, *Minimizing finite sums with the stochastic average gradient*, Mathematical Programming **162** (2017), no. 1, 83–112.
- Sharan Vaswani, Benjamin Dubois-Taine, and Reza Babanezhad, *Towards noise-adaptive, problem-adaptive stochastic gradient descent*, arXiv preprint arXiv:2110.11442 (2021).
- Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien, *Painless stochastic gradient: Interpolation, line-search, and convergence rates*, Advances in neural information processing systems **32** (2019).