By law of total expectation, we have on the overall result of this stochastic process.

$$\mathbb{E}[f(x_{k+1})] \le \mathbb{E}[f(x_k)] - \underbrace{(\alpha_k - \frac{L\alpha_k^2}{2})}_{(1-\frac{L\alpha_i}{2})}\mathbb{E}\|\nabla f(x_k)\|^2 + \frac{L\alpha_k^2\sigma^2}{2}$$

Induction on this ensures

$$\min f(x) \le \mathbb{E}[f(x_{k+1})] \le \underbrace{\mathbb{E}[f(x_0)]}_{= f(x_0)} - \sum_{i=0}^{k}(\alpha_i - \frac{L\alpha_i^2}{2})\,\mathbb{E}\|\nabla f(x_i)\|^2 + \sum_{i=0}^{k}\frac{L\alpha_i^2\sigma^2}{2}.$$

$$\overset{\text{Linearity}}{\Rightarrow} \mathbb{E}\Bigg[\underbrace{\sum_i \frac{\alpha_i(1-\frac{L\alpha_i}{2})\|\nabla f(x_i)\|^2}{\sum_j \alpha_j(1-\frac{L\alpha_j}{2})}}_{\text{a weighted average of the gradient norm squared}}\Bigg] \le \frac{f(x_0) - \min f + \sum \frac{L\alpha_i^2\sigma^2}{2}}{\sum \alpha_i(1-\frac{L\alpha_i}{2})}$$

$$\Rightarrow \mathbb{E}\Big[\min_{i\le k}\|\nabla f(x_i)\|^2\Big] \le \frac{f(x_0)-\min f + \sum \frac{L\alpha_i^2\sigma^2}{2}}{\sum \alpha_i(1-\frac{L\alpha_i}{2})}. \quad \square$$

Next time this is a $O(\frac{1}{\sqrt{k}})$ rate.

Pick $\alpha_k = \frac{1}{L\sqrt{T+1}}$, $\Rightarrow 1 - \frac{L\alpha_i}{2} \ge \frac{1}{2}$, then RHS $\le \dfrac{f(x_0) - \min f + \frac{\sigma^2}{2L}}{\frac{1}{2}(T+1)\frac{1}{L\sqrt{T+1}}}$

$$= \frac{L(f(x_0) - \min f) + \sigma^2/2}{\frac{1}{2}\sqrt{T+1}}.$$

$$\Rightarrow \mathbb{E}\left[\min_{i \le K} \|\nabla f(x_i)\|\right] \le O\left(\frac{1}{K^{1/4}}\right).$$

Hope for better guarantees under convexity.

3. Convex Guarantees

For ease, lets look at $\alpha_K = \alpha \le 1/L$.

<u>Theorem</u> In addition to the previous theorem, suppose f is convex. Then

$$\mathbb{E}\left[\min_{i \le K+1} \{f(x_i) - f(x^*)\}\right] \le \frac{\|x_0 - x^*\|^2}{2\alpha(K+1)} + \alpha\sigma^2$$

In particular, $\alpha = 1/\sqrt{K+1}$, for $K \ge L^2$

$$\Rightarrow \text{RHS} = \frac{\|x_0 - x^*\|^2 + 2\sigma^2}{2\sqrt{K+1}}.$$

Proof. By previous proof, we have a "stochastic" descent lemma

$$\mathbb{E}[f(x_{K+1}) | x_K] \le \underbrace{f(x_K)}_{\le f(x^*) - \nabla f(x_K)^T(x^* - x_K)} - \frac{\alpha}{2} \underbrace{\|\nabla f(x_K)\|^2}_{\ge \mathbb{E}[\|g(x_K)\|^2 | x_K] - \sigma^2} + \frac{\alpha\sigma^2}{2}.$$

$$\le f(x^*) - \nabla f(x_K)^T(x^* - x_K) - \frac{\alpha}{2}\mathbb{E}[\|g(x_K)\|^2 | x_K] + \alpha\sigma^2.$$

by unbiased and using Linearity.

$$= f(x^*) - \mathbb{E}\left[\alpha g(x_k)^T(x^* - x_k) + \frac{\alpha}{2}\|g(x_k)\|^2 \,\middle|\, x_k\right] + \alpha\sigma^2$$

by def of x_{k+1}

$$= f(x^*) - \mathbb{E}\left[g(x_k)^T(x^* - x_k) + \frac{1}{2\alpha}\|x_{k+1} - x_k\|^2 \,\middle|\, x_k\right] + \alpha\sigma^2$$

$$= f(x^*) - \mathbb{E}\left[\frac{1}{2\alpha}\left(\underbrace{\|x_{k+1} - x^*\|^2}_{=\ \|(x_k - x^*) - \alpha g(x_k)\|^2} - \|x_k - x^*\|^2\right) \,\middle|\, x_k\right] + \alpha\sigma^2$$

Law of total expectation:

$$\mathbb{E}[f(x_{k+1}) - f(x^*)] \leq \mathbb{E}\left[\frac{1}{2\alpha}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)\right] + \alpha\sigma^2$$

Summing these up and dividing by K gives

$$\frac{1}{K}\sum \mathbb{E}[f(x_{k+1}) - f(x^*)] \leq \frac{\|x_0 - x^*\|^2}{2\alpha \cdot K} + \alpha\sigma^2. \qquad \square$$

Note nonsmooth (but deterministic) also $\frac{1}{\sqrt{K}}$ rate.

Show nonsmooth stochastic opt has $\frac{1}{\sqrt{K}}$ rate.

HW 4 Q1.

(there want $g(x)$ s.t. $\mathbb{E}[g(x)] \in \partial f(x)$)

4. Improvements

Acceleration? Not really.

$$O\left(\frac{L\|x_0 - x^*\|^2}{k^2} + \frac{\sigma^2}{\sqrt{k}}\right) \quad \text{best possible.}$$

Randomized Coordinate "Descent"

Fix $g(x) = d \cdot \frac{\partial f}{\partial x_i}(x) \cdot e_i$, $i \sim$ Uniformly $\{1, \dots, d\}$

$$x_{k+1} = x_k - \frac{1}{Ld} g(x_k).$$

This is a descent method...

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

$$\cancel{= f(x_k) + \nabla f(x_k)^T}$$

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

$$= f(x_k) - \frac{1}{Ld} \nabla f(x_k)^T \left(d \frac{\partial f}{\partial x_i}(x_k) e_i\right) + \frac{1}{2Ld^2}\left(d \frac{\partial f}{\partial x_i}(x_k)\right)^2$$

$$= f(x_k) - \frac{1}{L}\left(\frac{\partial f}{\partial x_i}(x_k)\right)^2 + \frac{1}{2L}\left(\frac{\partial f}{\partial x_i}(x_k)\right)^2$$

$$= f(x_k) - \frac{1}{2L}\left(\frac{\partial f}{\partial x_i}(x_k)\right)^2.$$

$$\mathbb{E}[f(x_{k+1})] \le \mathbb{E}[f(x_k)] - \frac{1}{2L}\mathbb{E}\left[\left(\frac{\partial f}{\partial x_i}(x_k)\right)^2\right].$$

$$= \mathbb{E}[f(x_k)] - \frac{1}{2L}\cdot\frac{1}{d}\mathbb{E}\|\nabla f(x_k)\|^2$$

Iteratively apply that

$$\Rightarrow \quad \mathbb{E}[f(x_{k+1})] \le \mathbb{E}[f(x_0)] - \frac{1}{2Ld}\sum \mathbb{E}\left[\|\nabla f(x_k)\|^2\right]$$

$$\Rightarrow \mathbb{E}\left[\min_{k\le T}\|\nabla f(x_k)\|^2\right] \le \frac{2L\cdot d\cdot(f(x_0)-\min f)}{T}.$$

Same order of magnitude as our classic nonconvex, smooth GD rate.

Select i with largest $\frac{\partial f}{\partial x_i}(x_k)$ (Gauss-Southwell Rule)

[ICML'15, Nutini et al.]

Cyclic Coordinate Descent.

(similar guarantees) [Beck, Tetruashvili 2015, SIOPT]

Stochastic Variance Reduced Gradient Method (SVRG)

Family: $\left\{ \begin{array}{l} \text{SAGA} \\ \text{SAGA} \\ \text{SDCA} \\ \vdots \end{array} \right\}$

$\tilde{x}_0 \leftarrow x_0$

Compute full gradient $\nabla f(\tilde{x})$

outer loop

for $i = 0, \ldots,$

$y_0 = \tilde{x}_i$

inner loop

for $j = 0, \ldots, 2d$

$y_{j+1} = y_j - \alpha_k \, g(y_j)$

SVRG: $\nabla f(\tilde{x}_i) + \nabla f_i(y_j) - \nabla f_i(\tilde{x}_i)$

$i \sim$ Uniformly

end for

$\tilde{x}_{i+1} = \text{average}(y_0, \ldots y_{2d})$.

Compute full gradient $\nabla f(\tilde{x}_{i+1})$

end for

Storage: Each step $\left\{ \begin{array}{l} \nabla f(\tilde{x}_i) \\ \nabla f_i(y_j) \\ \nabla f_i(\tilde{x}_i) \end{array} \right.$

$\left\{ \begin{array}{l} y_j \\ \tilde{x}_i \end{array} \right.$

[Johnson Zhang 2013]

Theorem If f is L-smooth μ-strongly convex, ~~with~~ SVRG converges linearly.

[For comparison, stochastic GD, $O(1/k)$ rate]