

Chapter 12 Stochastic Optimization

- Formulation $\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) := \mathbb{E}_{\xi}[f(\mathbf{x}, \xi)]$ or special case of $\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$.
 - $\xi \in \Xi \subset \mathbb{R}^m$ is a random vector with distribution P
 - usually P unknown but can be sampled through data, F and ∇F usually hard to compute even if P given.
- Algorithm
 - $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, \xi_t)$ where $\xi_t \stackrel{iid}{\sim} P(\xi)$.
 - $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \nabla f_{i_t}(\mathbf{x}_t)$ where $i_t \in [n]$ is uniformly sampled.
- We need $\gamma_t \rightarrow 0$ to achieve convergence.
- To reduce noise, we can use mini-batch $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \cdot \frac{1}{b} \sum_{j \in J, |J|=b} \nabla f(\mathbf{x}_t, \xi_j)$ or SGD with iterate averaging $\bar{\mathbf{x}}_t = \frac{1}{t} \sum_{\tau=1}^t \mathbf{x}_\tau$.

Convergence for convex functions

- SGD is a special case of stochastic mirror descent, $\mathbf{x}_{t+1} = \arg \min_{\mathbf{x} \in X} \{V_\omega(\mathbf{x}, \mathbf{x}_t) + \langle \gamma_t G(\mathbf{x}_t, \xi_t), \mathbf{x} \rangle\}$.
 - $\mathbb{E}[G(\mathbf{x}, \xi)] \in \partial F(\mathbf{x})$ and we assume $\mathbb{E}[\|G(\mathbf{x}, \xi)\|_*^2] \leq M^2$.
- Theorem 12.4 F convex, then SMD gives $\mathbb{E}[F(\hat{\mathbf{x}}_T) - F(\mathbf{x}^*)] \leq \frac{R^2 + \frac{M^2}{2} \sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t}$, where $R^2 = \max_{\mathbf{x} \in X} V_\omega(\mathbf{x}, \mathbf{x}_1)$ and $\hat{\mathbf{x}}_T = \frac{\sum_{t=1}^T \gamma_t \mathbf{x}_t}{\sum_{t=1}^T \gamma_t}$.
 - Proof
 - The descent lemma gives $\gamma_t (f(\mathbf{x}_t, \xi_t) - f(\mathbf{x}^*, \xi_t)) \leq V_\omega(\mathbf{x}^*, \mathbf{x}_t) - V_\omega(\mathbf{x}^*, \mathbf{x}_{t+1}) + \frac{\gamma_t^2}{2} \|G(\mathbf{x}_t, \xi_t)\|_{a^*}^2$
 - In the proof of this descent lemma, no information of \mathbf{x}^* is used, so we can use this on $\mathbf{x}^* := \arg \min_x F(x)$.
 - Taking expectation w.r.t. ξ_t and condition on \mathbf{x}_t we get (\mathbf{x}_t is still stochastic.)
 - $\gamma_t (F(\mathbf{x}_t) - F(\mathbf{x}^*)) \leq V_\omega(\mathbf{x}^*, \mathbf{x}_t) - \mathbb{E}[V_\omega(\mathbf{x}^*, \mathbf{x}_{t+1}) | \mathbf{x}_t] + \frac{\gamma_t^2}{2} M^2$
 - Or equivalently under 2-norm we can get with only convexity of F ,
 - $2\gamma_t (F(\mathbf{x}_t) - F(\mathbf{x}^*)) \leq \|\mathbf{x}^* - \mathbf{x}_t\|_2^2 - \mathbb{E}[\|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2 | \mathbf{x}_t] + \gamma_t^2 \|G(\mathbf{x}_t, \xi_t)\|_2^2$
 - Further taking expectation w.r.t \mathbf{x}_t and condition on \mathbf{x}_{t-1} and we get
 - $\gamma_t \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*) | \mathbf{x}_{t-1}] \leq \mathbb{E}[V_\omega(\mathbf{x}^*, \mathbf{x}_t) - V_\omega(\mathbf{x}^*, \mathbf{x}_{t+1}) | \mathbf{x}_{t-1}] + \frac{\gamma_t^2}{2} M^2$
 - add w.r.t. $t-1$ case and we get

$$\gamma_{t-1} (F(\mathbf{x}_{t-1}) - F(\mathbf{x}^*)) + \gamma_t \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*) | \mathbf{x}_{t-1}] \leq V_\omega(\mathbf{x}^*, \mathbf{x}_{t-1}) - \mathbb{E}[V_\omega(\mathbf{x}^*, \mathbf{x}_{t+1}) | \mathbf{x}_t] + (\gamma_{t-1}^2 + \gamma_t^2) M^2 / 2$$
 - do this expectation and summation through $t \in [1 : T]$ and we get
 - $\sum_{t=1}^T \gamma_t \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*) | \mathbf{x}_1] \leq V_\omega(\mathbf{x}^*, \mathbf{x}_1) - \mathbb{E}[V_\omega(\mathbf{x}^*, \mathbf{x}_{T+1}) | \mathbf{x}_t] + \sum_{t=1}^T \gamma_t^2 M^2 / 2$
 - then the conclusion is straightforward.
 - If we set $\gamma_t \equiv \frac{R}{M\sqrt{T}}$, then $\mathbb{E}[F(\hat{\mathbf{x}}_T) - F(\mathbf{x}^*)] = O\left(\frac{BR}{\sqrt{T}}\right)$, this means $O(1/\epsilon^2)$ of sample complexity.

Convergence for strongly convex functions

- Descent Lemma assume f is μ -strongly convex and $\mathbb{E}[\|\nabla f(\mathbf{x}, \xi)\|_2^2] \leq B^2$, then

$$\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2] \leq (1 - 2\mu\gamma_t) \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] + \gamma_t^2 B^2$$
 - Proof
 - $\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 \leq \|\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, \xi_t) - \mathbf{x}^*\|_2^2 = \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\gamma_t \langle \nabla f(\mathbf{x}_t, \xi_t), \mathbf{x}_t - \mathbf{x}^* \rangle + \gamma_t^2 \|\nabla f(\mathbf{x}_t, \xi_t)\|_2^2$
 - Taking expectation over ξ_t and we get
 - $\mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2 | \mathbf{x}_t] \leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\gamma_t \langle \nabla F(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle + \gamma_t^2 B^2$
 - Strong convexity of F means coercivity, $\langle \nabla F(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle = \langle \nabla F(\mathbf{x}_t) - F(\mathbf{x}^*), \mathbf{x}_t - \mathbf{x}^* \rangle \geq \mu \|\mathbf{x}_t - \mathbf{x}^*\|_2^2$
 - plug this in and we get the result.
- Theorem 12.3 Assume $F(\mathbf{x})$ is μ -strongly convex, and $\exists B > 0$, s.t. $\forall \mathbf{x} \in X, \mathbb{E}[\|\nabla f(\mathbf{x}, \xi)\|_2^2] \leq B^2$, then SGD with $\gamma_t = \gamma/t$ at iteration t where $\gamma > 1/2\mu$ satisfies $\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2] \leq \frac{C(\gamma)}{t}$ where $C(\gamma) = \max\left\{\frac{\gamma^2 B^2}{2\mu\gamma-1}, \|\mathbf{x}_1 - \mathbf{x}^*\|_2^2\right\}$.
 - Proof
 - By decsetn lemma than take expectation over all t , denote $\varepsilon_t := \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 | \mathbf{x}_1]$, we have
 - $\varepsilon_{t+1} \leq (1 - 2\mu\gamma/t) \varepsilon_t + \gamma^2 B^2 t^2$.
 - We use induction and assume $\exists C(\gamma)$ s.t. $\forall \tau \leq t, \varepsilon_\tau \leq C(\gamma)/\tau$.
 - To make $\varepsilon_{t+1} \leq C(\gamma)/(t+1)$, we need (denote $b := 2\mu\gamma > 1$ and $a := B^2\gamma^2$)
 - $\varepsilon_{t+1} \leq \frac{C(\gamma)}{t} \left(1 - \frac{b}{t}\right) + \frac{a}{t^2} = \frac{C(\gamma)}{t+1} \left(1 + \frac{1}{t}\right) \left(1 - \frac{b}{t}\right) + \frac{a}{t^2} = \frac{C(\gamma)}{t+1} - \frac{C(\gamma)}{t(t+1)} \left(b - 1 + \frac{b}{t}\right) + \frac{a}{t^2}$
 - We need the residual term $\frac{C(\gamma)}{t(t+1)} \left(b - 1 + \frac{b}{t}\right) - \frac{a}{t^2} > 0$

- equivalently $C(\gamma) \geq \frac{1}{b-1+\frac{b}{t}}(1 + \frac{1}{t})a = \frac{a}{b-1} \frac{1+t}{b/(b-1)+t}$.
 - Since $b \geq 1$, $b/(b-1) > 1$, $\frac{1+t}{b/(b-1)+t}$ increase w.r.t t , so $C(\gamma) \geq \max_t \frac{1}{b-1+\frac{b}{t}}(1 + \frac{1}{t})a = \frac{a}{b-1} \frac{1+\infty}{b/(b-1)+\infty} = \frac{a}{b-1} = \frac{\gamma^2 B^2}{2\mu\gamma-1}$
 - To ensure induction succeed, we also need $C(\gamma) \geq \varepsilon_1 = \|\mathbf{x}_1 - \mathbf{x}_*\|_2^2$.
 - sample complexity is $O(1/\epsilon)$.
 - If F also L -smooth, we have $\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] = O\left(\frac{L \cdot C(\gamma)}{t}\right)$.
- Example SGD over $\min_x F(x) := \frac{1}{2}\mathbb{E}_{\xi \sim N(0,1)}[(x - \xi)^2]$ with $\gamma_t = 1/t$ gives $\mathbb{E}[|x_{t+1} - x^*|^2] = \frac{1}{t}$, since $t x_{t+1} = (t-1)x_t + \xi_t$.
- Lemma (Boundedness of Stochastic Gradients, Ex5.1 5.2) $F(\mathbf{x}) = \mathbb{E}[f(\mathbf{x}, \xi)]$ and f is convex and L -smooth, define $\mathbf{x}^* = \operatorname{argmin}_{\mathbf{x}} F(\mathbf{x})$, then
 - $\mathbb{E}[\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x}^*, \xi)\|_2^2] \leq 2L[F(\mathbf{x}) - F(\mathbf{x}^*)]$ and $\mathbb{E}[\|\nabla f(\mathbf{x}, \xi)\|_2^2] \leq 4L[F(\mathbf{x}) - F(\mathbf{x}^*)] + 2\mathbb{E}[\|\nabla f(\mathbf{x}^*, \xi)\|_2^2]$
 - Proof
 - first
 - Define $f_y(\mathbf{x}) := f(\mathbf{x}, \xi) - \nabla f(\mathbf{y}, \xi)^\top (\mathbf{x} - \mathbf{y})$, then $\mathbf{x} = \mathbf{y}$ is the minima for f_y ,
 - by Lemma B in chap3, the quadratic bound for smooth function
 - $f(\mathbf{x}^*, \xi) - \nabla f(\mathbf{y}, \xi)^\top (\mathbf{x}^* - \mathbf{y}) - f(\mathbf{y}, \xi) = f_y(\mathbf{x}^*) - f_y(\mathbf{y}) \geq \frac{1}{2L} \|\nabla f_y(\mathbf{x}^*)\|_2^2 = \frac{1}{2L} \|\nabla f(\mathbf{x}^*) - \nabla f(\mathbf{x}_y)\|_2^2$.
 - Taking expectation over ξ and $\nabla F(\mathbf{x}^*) = 0$, we get
 - $F(\mathbf{x}^*) - F(\mathbf{y}) \geq \frac{1}{2L} \mathbb{E}[\|\nabla f(\mathbf{x}^*) - \nabla f(\mathbf{x}_y)\|_2^2]$
 - Second
 - by $\|\mathbf{a} + \mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$
 - so $\|\nabla f(\mathbf{x}, \xi)\|_2^2 \leq 2\|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x}^*, \xi)\|_2^2 + 2\|\nabla f(\mathbf{x}^*, \xi)\|_2^2$.
- Lemma of my own (might be useful in exam) f convex and L -smooth, then $\forall r, f(\mathbf{x}) - f(\mathbf{y}) \geq \nabla f_y^\top (\mathbf{x} - \mathbf{y}) - (\nabla f_x - \nabla f_y)^\top r - L\|r\|_2^2/2$
 - Proof
 - By smoothness, $f(\mathbf{x} + r) \leq f(\mathbf{x}) + \nabla f_x^\top r + L\|r\|_2^2/2$
 - By convexity $f(\mathbf{x} + r) \geq f(\mathbf{y}) + \nabla f_y^\top (\mathbf{x} + r - \mathbf{y})$
 - combine them together and we get the result.
- Convergence of SGD under constant stepsize
 - Theorem 12.5 Assume F is μ -strongly convex and L -smooth, and also $\mathbb{E}[\|\nabla f(\mathbf{x}, \xi)\|_2^2] \leq \sigma^2 + c\|\nabla F(\mathbf{x})\|_2^2$, then SGD with constant step size $\gamma_t \equiv \gamma \leq \frac{1}{Lc}$ gives $\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] \leq \frac{\gamma L \sigma^2}{2\mu} + (1 - \gamma\mu)^{t-1} [F(\mathbf{x}_1) - F(\mathbf{x}^*)]$.
 - Proof
 - By Smoothness of F , we have $F(\mathbf{x}_{t+1, \xi_t}) - F(\mathbf{x}_t) \leq -\gamma \langle \nabla F(\mathbf{x}_t), \nabla f(\mathbf{x}_t, \xi_t) \rangle + \frac{L\gamma^2}{2} \|\nabla f(\mathbf{x}_t, \xi_t)\|_2^2$
 - Taking expectation over ξ_t , we get $\mathbb{E}[F(\mathbf{x}_{t+1}) | \mathbf{x}_t] - F(\mathbf{x}_t) \leq -\gamma \|\nabla F(\mathbf{x}_t)\|_2^2 + \frac{L\gamma^2}{2} (\sigma^2 + c\|\nabla F(\mathbf{x}_t)\|_2^2)$
 - By strong convexity we have $\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}^*)\| \geq 2\mu(F(\mathbf{x}_t) - F(\mathbf{x}^*))$, take half of this term into above and we get
 - $\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*) | \mathbf{x}_t] \leq (1 - \gamma\mu)(F(\mathbf{x}_t) - F(\mathbf{x}^*)) + \frac{L\gamma^2\sigma^2}{2} - (1 - L\gamma c)\frac{\gamma}{2} \|\nabla F(\mathbf{x}_t)\|_2^2$
 - Since $\gamma \geq 1/cL$, we get $\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \leq (1 - \gamma\mu)\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] + \frac{L\gamma^2\sigma^2}{2}$
 - $(1 - \gamma\mu)^{-T+1} \mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^*)] \leq \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] + \sum_{t=1}^{T-1} (1 - \gamma\mu)^{-t} \frac{L\gamma^2\sigma^2}{2}$, else is easy.
 - If we start from $\|\mathbf{x}_t - \mathbf{x}^*\|_2^2$ we can also get similar conclusion on it.
 - This means constant stepsize converges linearly to some neighborhood of the optimal solution.
 - If variance is bounded $\mathbb{E}[\|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|_2^2] \leq \sigma^2$, then the condition naturally holds with $c = 1$.
 - When $\mathbb{E}[\|\nabla f(\mathbf{x}, \xi)\|_2^2] \leq c\|\nabla F(\mathbf{x})\|_2^2$, the case is called *strong growth condition* with constant c . SGD then converge to global optimal.
 - When $F(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$, strong growth means *interpolation*, $\nabla f_i(\mathbf{x}^*) = 0, \forall i$.
 - Example: linear regression or over-parametrized neural network in the realizable case.
 - Lemma (PL -> strong growth) If f is L -smooth and F is μ -PL, then we have strong growth condition (I cannot prove assuming F is smooth).
 - Proof
 - By Lemma B in chp3 $f(\mathbf{x}, \xi) - f(\mathbf{x}^*, \xi) \geq \frac{1}{2L} \|\nabla f(\mathbf{x}, \xi)\|_2^2 + \text{Linear Term}$, taking expectation over ξ
 - $\mathbb{E}[\|\nabla f(\mathbf{x}, \xi)\|_2^2] \leq 2L(F(\mathbf{x}) - F(\mathbf{x}^*))$, then by PL of F , $F(\mathbf{x}) - F(\mathbf{x}^*) \leq \frac{1}{2\mu} \|\nabla F(\mathbf{x})\|_2^2$ we get strong growth condition with $c = L/\mu$.
 - Definition (Weak Growth) $F = \mathbb{E}f$ is L -smooth and have a minima \mathbf{x}^* , stochastic gradient satisfies the weak growth condition with

constant c if $\mathbb{E} [\|\nabla f(\mathbf{x}, \xi)\|_2^2] \leq 2cL [F(\mathbf{x}) - F(\mathbf{x}_*)]$.

- Lemma (Ex 75.1) If F convex, strong growth \rightarrow weak growth.
 - Proof By smoothness $\|\nabla F(\mathbf{x})\|_2^2 \leq 2L[F(\mathbf{x}) - F(\mathbf{x}^*)]$ (PS: I don't see why assume convex)
- Lemma (Ex 75.2) If F μ -strong convex, weak growth \rightarrow strong growth.
 - Proof By strong convexity $\|\nabla F(\mathbf{x})\|_2^2 \geq 2\mu[F(\mathbf{x}) - F(\mathbf{x}^*)]$

Convergence for nonconvex functions (handout 11)

- Theroem 12.8 If $\text{dom}(F) = X = \mathbb{R}^n$, F is L -smooth and $\mathbb{E} [\|\nabla f(\mathbf{x}, \xi) - \nabla F(\mathbf{x})\|_2^2] \leq \sigma^2$, then stepsize of $\gamma_t = \min \left\{ \frac{1}{L}, \frac{\gamma}{\sigma\sqrt{T}} \right\}$ gives $\mathbb{E} [\|\nabla F(\hat{\mathbf{x}}_T)\|_2^2] \leq \frac{\sigma}{\sqrt{T}} \left(\frac{2(F(\mathbf{x}_1) - F(\mathbf{x}_*))}{\gamma} + L\gamma \right)$, where $\hat{\mathbf{x}}_T$ is selected uniformly at random from $\{\mathbf{x}_1, \dots, \mathbf{x}_T\}$.
 - Proof
 - similar to privous proof, $\mathbb{E}[F(\mathbf{x}_{t+1})|\mathbf{x}_t] - F(\mathbf{x}_t) \leq -\gamma\|\nabla F(\mathbf{x}_t)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}\|\nabla f(\mathbf{x}_t, \xi_t)\|_2^2$.
 - By our assumption, we have $\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}_t)] \leq -\frac{\gamma_t}{2}\mathbb{E}\|\nabla F(\mathbf{x}_t)\|^2 + \frac{L\sigma^2\gamma_t^2}{2}$, sum over $t \in [1 : T]$ and we get the result.
 - This means in non-convex setting, finding ϵ -stationary point $\mathbb{E}[\|\nabla F(\bar{\mathbf{x}})\|] \leq \epsilon$ requirese at most $O(1/\epsilon^4)$ iteration/data point.

Lower Bound

- Sample complexity of $O(1/\epsilon^2)$ and $O(1/\epsilon)$ for convex / strongly convex cannot be improved.
- Definition (Stochastic Oracle) Given input \mathbf{x} , stochastic oracle returns $G(\mathbf{x}, \xi)$ such that $\mathbb{E}[G(\mathbf{x}, \xi)] \in \partial F(\mathbf{x})$ and $\mathbb{E} [\|G(\mathbf{x}, \xi)\|_p^2] \leq M^2$, $M > 0$ and $p \in [1, \infty]$.
- Theorem (Agarwal et al., 2012) Let $X = B_\infty(r)$ be a ℓ_∞ ball in \mathbb{R}^d .
 - i. $\exists c_0 > 0$, convex function f with $|f(x) - f(y)| < M\|x - y\|_\infty$, for any algorithm making T stochastic oracles with $1 \leq p \leq 2$,
 - $\mathbb{E}[f(\mathbf{x}_T) - f(\mathbf{x}^*)] \geq \min \left\{ c_0Mr\sqrt{\frac{d}{T}}, \frac{Mr}{144} \right\}$
 - ii. $\exists c_1, c_2 > 0$, μ -strongly convex function f , for any algorithm making T stochastic oracles with $p = 1$,
 - $\mathbb{E}[f(\mathbf{x}_T) - f(\mathbf{x}^*)] \geq \min \left\{ c_1\frac{M^2}{\mu^2 T}, c_2Mr\sqrt{\frac{d}{T}}, \frac{M^2}{1152\mu^2 d}, \frac{Mr}{144} \right\}$.

Adaptive Stochastic Gradient Methods

- Generic Adaptive Scheme
 - $\mathbf{g}_t = \nabla f(\mathbf{x}_t, \xi_t)$
 - $\mathbf{m}_t = \phi_t(\mathbf{g}_1, \dots, \mathbf{g}_t)$
 - $V_t = \psi_t(\mathbf{g}_1, \dots, \mathbf{g}_t)$
 - $\hat{\mathbf{x}}_t = \mathbf{x}_t - \alpha_t V_t^{-1/2} \mathbf{m}_t$
 - $\mathbf{x}_{t+1} = \underset{\mathbf{x} \in X}{\operatorname{argmin}} \left\{ (\mathbf{x} - \hat{\mathbf{x}}_t)^T V_t^{1/2} (\mathbf{x} - \hat{\mathbf{x}}_t) \right\}$
- SGD $\phi_t(\mathbf{g}_1, \dots, \mathbf{g}_t) = \mathbf{g}_t$, $\psi_t(\mathbf{g}_1, \dots, \mathbf{g}_t) = \mathbb{I}$
- AdaGrad $\phi_t(\mathbf{g}_1, \dots, \mathbf{g}_t) = \mathbf{g}_t$, $\psi_t(\mathbf{g}_1, \dots, \mathbf{g}_t) = \operatorname{diag} \left(\sum_{\tau=1}^t \mathbf{g}_\tau^2 \right) / t$
 - can be viewed as Mirror descent with Bregman Divergence of $\omega_t(\mathbf{x}) = \frac{1}{2} \mathbf{x}_t^T H_t \mathbf{x}$, where $H_t = \epsilon \mathbf{I} + \left[\sum_{t=1}^t \mathbf{g}_t \mathbf{g}_t^T \right]^{\frac{1}{2}}$.
- RMSProp $\phi_t(\mathbf{g}_1, \dots, \mathbf{g}_t) = \mathbf{g}_t$, $\psi_t(\mathbf{g}_1, \dots, \mathbf{g}_t) = (1 - \beta_2) \operatorname{diag} \left(\sum_{\tau=1}^t \beta_2^{t-\tau} \mathbf{g}_\tau^2 \right)$
 - momentum on gradient norm term to slow down the decay of learning rate.
- Adam $\phi_t(\mathbf{g}_1, \dots, \mathbf{g}_t) = (1 - \beta_1) \sum_{\tau=1}^t \beta_1^{t-\tau} \mathbf{g}_\tau$, $\psi_t(\mathbf{g}_1, \dots, \mathbf{g}_t) = (1 - \beta_2) \operatorname{diag} \left(\sum_{\tau=1}^t \beta_2^{t-\tau} \mathbf{g}_\tau^2 \right)$,
 - equivalently $\mathbf{m}_t = \beta_1 \mathbf{m}_{t-1} + (1 - \beta_1) \mathbf{g}_t$, $V_t = \beta_2 V_{t-1} + (1 - \beta_2) \operatorname{diag}(\mathbf{g}_t^2)$.
 - momentum on both gradient and gradient norm term.
- Adaptive Methods
 - Less sensitive to parameter tuning and adapt to sparse gradients.
 - Out perform SGD for NLP tasks, training GANs, deep RL, etc., but are less effective in CV tasks.
 - Tend to overfit and generalize worse than their non-adaptive counterparts.

- Often display faster initial progress on the training set, but their performance quickly plateaus on the testing set.

- What we know in theory

- SGD with momentum has no acceleration even for some convex quadratic functions.
- For convex problems, Adagrad does converge, but RMSProp and Adam may not when $\beta_1 < \sqrt{\beta_2}$.

- Example of non-convergence of Adam

- $X = [-1, 1]$, $f(x, \xi) = \begin{cases} Cx, & \text{if } \xi = 1 \\ -x, & \text{if } \xi = 0 \end{cases}$, $P(\xi = 1) = p = \frac{1+\delta}{C+1}$
- $F(x) = \mathbb{E}[f(x, \xi)] = \delta x$ and $x^* = -1$.
- update rule gives $x_{t+1} = x_t - \gamma_0 \Delta_t$ with $\Delta_t = \frac{\alpha m_t + (1-\alpha)g_t}{\sqrt{\beta v_t + (1-\beta)g_t^2}}$
- For C large enough, one can show that $\mathbb{E}[\Delta_t] \leq 0$.

- A fix: *AMSGrad*, can prove convergence for many convex case.

- changes: $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$, $\hat{v}_t = \max(\hat{v}_{t-1}, v_t)$ and $\hat{V}_t = \text{diag}(\hat{v}_t)$.
- Idea: if $g_t \ll m_t$ then $v_t < v_{t+1}$, this might increase step size.

Variance reduce methods

- Idea $\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] \leq \frac{\gamma L \sigma^2}{2\mu} + (1 - \mu\gamma)^{t-1} (F(\mathbf{x}_1) - F(\mathbf{x}^*))$, if σ^2 can be reduced → better upper bound.
- Mini-Batching $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \frac{1}{b} \sum_{i=1}^b \nabla f(\mathbf{x}_t, \xi_{t,i})$, $\sigma^2 \leftarrow \sigma^2/b$ but at cost of computation. sample complexity remains the same.
- Importance sampling since $\xi \sim P$, we can change to another distri Q by $G(\mathbf{x}_t, \xi_t) \Rightarrow G(\mathbf{x}_t, \eta_t) \frac{P(\eta_t)}{Q(\eta_t)}$, and variance may be smaller.
- Momentum $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \hat{\mathbf{m}}_t$ where $\hat{\mathbf{m}}_t = c \cdot \sum_{\tau=1}^t \alpha^{t-\tau} \nabla f_{i_\tau}(\mathbf{x}_\tau)$
- Key Idea of Modern Variance Reduction
 - we want to estimate $\theta = \mathbb{E}[X]$, but estimator is $\hat{\theta} := X - Y$, where $\mathbb{E}[Y] = 0$.
 - $\mathbb{V}[X - Y] < \mathbb{V}[X]$ if $\text{Cov}(X, Y) > 0$
- Point Estimator $\hat{\theta}_\alpha = \alpha(X - Y) + \mathbb{E}[Y]$, $\mathbb{E}[\hat{\theta}_\alpha] = \alpha \mathbb{E}[X] + (1 - \alpha) \mathbb{E}[Y]$ and $\mathbb{V}[\hat{\theta}_\alpha] = \alpha^2 (\mathbb{V}[X] + \mathbb{V}[Y] - 2 \text{Cov}[X, Y])$.
 - $\alpha = 0 \rightarrow 1$, highly bias and no variance → unbiased but largest variance.
 - If $\text{Cov}[X, Y]$ is sufficiently large, then $\text{Var}[\hat{\theta}_\alpha] < \text{Var}[X]$
 - Idea $\mathbf{g}_t := \alpha(\nabla f_{i_t}(\mathbf{x}_t) - Y) + \mathbb{E}[Y]$ s.t. $\mathbb{E}[\|\mathbf{g}_t - \nabla F(\mathbf{x}_t)\|^2] \rightarrow 0$, as $t \rightarrow \infty$.
- Choice 1 $Y = \nabla f_{i_t}(\mathbf{x}^*)$, $\mathbb{E}[Y] = 0$, unrealistic but conceptually useful.
- Choice 2 $Y = \nabla f_{i_t}(\bar{\mathbf{x}}_{i_t})$, where $\bar{\mathbf{x}}_{i_t}$ is the last point for which we evaluated $\nabla f_i(\bar{\mathbf{x}}_i)$. $\mathbb{E}[Y] = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\bar{\mathbf{x}}_i)$, but requires storage of $\{\bar{\mathbf{x}}_i\}_{i=1}^n$ or $\{\nabla f_i(\bar{\mathbf{x}}_i)\}_{i=1}^n$
- Choice 3 $Y = \nabla f_{i_t}(\tilde{\mathbf{x}})$, where $\tilde{\mathbf{x}}$ is some fixed reference point. $\mathbb{E}[Y] = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{\mathbf{x}})$, and requires computing full gradient.

Stochastic Variance-Reduced Algorithms

Stochastic Average Gradient (SAG) ($\alpha = \frac{1}{n}$, $Y = \mathbf{v}_{i_t}$)

- $\mathbf{g}_t = \frac{1}{n}(\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{v}_{i_t}) + \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i$ where \mathbf{v}_i is the past gradient $\mathbf{v}_i^t = \begin{cases} \nabla f_{i_t}(\mathbf{x}_t), & \text{if } i = i_t \\ \mathbf{v}_i^{t-1}, & \text{if } i \neq i_t \end{cases}$
- Equivalently $\mathbf{g}_t = \mathbf{g}_{t-1} - \frac{1}{n} \mathbf{v}_{i_t}^{t-1} + \frac{1}{n} \nabla f_{i_t}(\mathbf{x}_t)$, or the update rule $\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\gamma}{n} \sum_{i=1}^n \mathbf{v}_i^t$.
- Same per-iteration cost as SGD but only additional memory cost of $O(nd)$.
- Theorem 12.10 (Schmidt et al. 2017, Linear Convergence) If F is μ -strongly convex and each f_i is L_i -smooth and convex. Setting $\gamma = 1/(16L_{\max})$ where $L_{\max} := \max_{i \in [n]} \{L_i\}$, SAG satisfies that $\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] \leq C \cdot \left(1 - \min\left\{\frac{1}{8n}, \frac{\mu}{16L_{\max}}\right\}\right)^t$.
- Full GD needs $O(\kappa \ln(\frac{1}{\epsilon}))$ iteration and $O(n)$ computation per-iter, so the total computation cost is $O(n\kappa \ln(\frac{1}{\epsilon}))$, where $\kappa = \overline{L_i}/\mu$.
- SAG needs $O((n + \kappa_{\max}) \ln(\frac{1}{\epsilon}))$ iteration and $O(1)$ computation per-iter, $O(n + \kappa) \ll O(n\kappa)$ may be true if n, κ are large.

SAGA ($\alpha = 1$, $Y = \mathbf{v}_{i_t}$, improved ver. of SAG)

- $\mathbf{g}_t = (\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{v}_{i_t}) + \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i$
- $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma [(\nabla f_{i_t}(\mathbf{x}_t) - \mathbf{v}_{i_t}^{t-1}) + \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i^{t-1}]$

- SAGA is unbiased, while SAG is biased, same $O(nd)$ memory cost as SAG, but proof much simpler.

Stochastic Variance Reduced Gradient (SVRG) ($\alpha = 1, Y = \nabla f_{i_t}(\tilde{\mathbf{x}})$)

- $\mathbf{g}_t = \nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})$
- The induction is that $\mathbb{E} [\|\mathbf{g}_t - \nabla F(\mathbf{x}_t)\|^2] \leq \mathbb{E} [\|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}})\|^2] \leq L_{\max}^2 \|\mathbf{x}_t - \tilde{\mathbf{x}}\|^2$
 - closer $\tilde{\mathbf{x}}$ to \mathbf{x}_t , smaller the variance, so update $\tilde{\mathbf{x}}$ in outer loop.
- Algorithm (Two-loop structure)
 - For $s = 1, 2, \dots$ do (outer loop)
 - Set $\tilde{\mathbf{x}} = \tilde{\mathbf{x}}^{s-1}$ and compute $\nabla F(\tilde{\mathbf{x}}) = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\tilde{\mathbf{x}})$ (n grad computation)
 - Initialize $\mathbf{x}_0 = \tilde{\mathbf{x}}$
 - For $t = 0, 1, \dots, m-1$ do (inner loop) ($2m$ grad computation)
 - Randomly pick $i_t \in \{1 : n\}$
 - Update $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta (\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}}))$
 - End For
 - Update $\tilde{\mathbf{x}}^s = \frac{1}{m} \sum_{t=0}^{m-1} \mathbf{x}_t$
 - End for
- Total of $O(n + 2m)$ component gradient evaluations at each outer epoch.
- Pros: no need to store past gradients or past iterates
- Cons: More parameter tuning, two gradient computation per iteration.
- Lemma 12.12 (Exercise 7.1) $f_i(\mathbf{x})$ is convex and L -smooth, then $\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^*)\|_2^2 \leq 2L_{\max} (F(\mathbf{x}) - F(\mathbf{x}^*))$
 - Proof is same as Exercise 6.1.
- Lemma A Denote $\mathbf{g}_t := \nabla f_{i_t}(\mathbf{x}^t) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})$ then $\mathbb{E} [\|\mathbf{g}_t\|_2^2] \leq 4L_{\max} [F(\mathbf{x}_t) - F(\mathbf{x}^*) + F(\tilde{\mathbf{x}}) - F(\mathbf{x}^*)]$.
 - Proof
 - Define $\mathbf{h}_i(\mathbf{x}) := f_i(\mathbf{x}) - \nabla f_i(\tilde{\mathbf{x}})^\top \mathbf{x} + \nabla F(\tilde{\mathbf{x}})^\top \mathbf{x}$, one can check that $H(\mathbf{x}) := \mathbb{E}[\mathbf{h}_i(\mathbf{x})] = F(\mathbf{x})$
 - By the conclusion in Exercise 6.2, we have $\mathbb{E}[\|\nabla \mathbf{h}_i(\mathbf{x}_t)\|_2^2] \leq 4L_{\max}[F(\mathbf{x}_t) - F(\mathbf{x}^*)] + 2\mathbb{E}[\|\nabla \mathbf{h}_i(\mathbf{x}^*)\|_2^2]$
 - By definition $\nabla \mathbf{h}_i(\mathbf{x}^*) := (\nabla f_i(\mathbf{x}^*) - \nabla f_i(\tilde{\mathbf{x}})) + (\nabla F(\tilde{\mathbf{x}}) - \nabla F(\mathbf{x}^*))$
 - Therefore $\mathbb{E}[\|\nabla \mathbf{h}_i(\mathbf{x}^*)\|_2^2] = \mathbb{E}[\|\nabla f_i(\mathbf{x}^*) - \nabla f_i(\tilde{\mathbf{x}})\|_2^2] + \|\nabla F(\tilde{\mathbf{x}}) - \nabla F(\mathbf{x}^*)\|_2^2 + 2\mathbb{E}[\nabla f_i(\mathbf{x}^*) - \nabla f_i(\tilde{\mathbf{x}})]^\top (\nabla F(\tilde{\mathbf{x}}) - \nabla F(\mathbf{x}^*))$
 - RHS = $\mathbb{E}[\|\nabla f_i(\mathbf{x}^*) - \nabla f_i(\tilde{\mathbf{x}})\|_2^2] - \|\nabla F(\tilde{\mathbf{x}}) - \nabla F(\mathbf{x}^*)\|_2^2 \leq \mathbb{E}[\|\nabla f_i(\mathbf{x}^*) - \nabla f_i(\tilde{\mathbf{x}})\|_2^2] \leq 2L_{\max}(F(\tilde{\mathbf{x}}) - F(\mathbf{x}^*))$, by Exercise 6.1.
 - Plug this in and we get the result.
- Theorem 12.11 (Johnson & Zhang, 2013, geometric convergence) Assume $f_i(\mathbf{x})$ is convex and L -smooth and $F(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x})$ is μ -strongly convex. Let $\mathbf{x}_* = \arg \min_{\mathbf{x}} F(\mathbf{x})$. Assume m is sufficiently large (and $\eta < 1/2L$), so that,

$$\rho = \frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1,$$
 - then we have geometric convergence in expectation $\mathbb{E} [F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}_*)] \leq \rho^s [F(\tilde{\mathbf{x}}^0) - F(\mathbf{x}_*)]$.
 - Proof
 - $\mathbb{E} [\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2] = \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta(\mathbf{x}_t - \mathbf{x}^*)^\top \mathbb{E}[\mathbf{g}_t] + \eta^2 \mathbb{E} [\|\mathbf{g}_t\|_2^2]$
 - By convexity and Lemma A RHS $\leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta(1-2L\eta)(F(\mathbf{x}_t) - F(\mathbf{x}^*)) + 4L\eta^2 [F(\mathbf{x}_0) - F(\mathbf{x}^*)]$
 - The above is for a single step of inner loop, we sum over $t \in [0 : m-1]$ and get
 - $\mathbb{E} [\|\mathbf{x}_m - \mathbf{x}^*\|_2^2] \leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - 2\eta(1-2L\eta) \sum_{t=0}^{m-1} \mathbb{E} [F(\mathbf{x}_t) - F(\mathbf{x}^*)] + 4mL\eta^2 [F(\mathbf{x}_0) - F(\mathbf{x}^*)]$
 - By convexity, we have $\sum_{t=0}^{m-1} (F(\mathbf{x}_t) - F(\mathbf{x}^*)) \geq m \left(F\left(\frac{1}{m} \sum_{t=0}^{m-1} \mathbf{x}_t\right) - F(\mathbf{x}^*) \right) = m(F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*))$
 - and by the condition of $\eta < 1/2L$, we have
 - $\mathbb{E} [\|\mathbf{x}_m - \mathbf{x}^*\|_2^2] \leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - 2\eta m(1-2L\eta) \mathbb{E} [F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*)] + 4mL\eta^2 [F(\mathbf{x}_0) - F(\mathbf{x}^*)]$
 - By definition $\mathbf{x}_0 = \tilde{\mathbf{x}}^{s-1}$, and by μ -strong convexity $\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \leq \frac{2}{\mu}(F(\mathbf{x}_0) - F(\mathbf{x}^*))$
 - $2\eta m(1-2L\eta) \mathbb{E} [F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*)] + \underbrace{\mathbb{E} [\|\mathbf{x}_m - \mathbf{x}^*\|_2^2]}_{\text{omitted}} \leq \left(\frac{2}{\mu} + 4mL\eta^2 \right) \mathbb{E} [F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}^*)]$
 - This means $\mathbb{E} [F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}_*)] \leq \left[\frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} \right] \mathbb{E} [F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}_*)]$. QED
 - Remark 12.13 Setting $\eta L = \theta$ gives $\rho = \frac{L}{\mu\theta(1-2\theta)m} + \frac{2\theta}{1-2\theta} = O\left(\frac{L}{\mu m} + \text{const.}\right)$, If further set $m = O(L/\mu)$, we get a constant ρ .
 - Then the number of epoch for ϵ optimal is $O\left(\log\left(\frac{1}{\epsilon}\right)\right)$. Total number of gradient computation required is $\mathcal{O}\left((m+n)\log\left(\frac{1}{\epsilon}\right)\right) = \mathcal{O}\left(\left(n + \frac{L}{\mu}\right)\log\left(\frac{1}{\epsilon}\right)\right)$
 - Remark
 - if use importance sampling $\mathbb{P}(i_t = i) = \frac{L_i}{\sum L_i}$, we get L_{avg} instead of L_{\max} .

- Incorporating acceleration, can improve to $\mathcal{O}((n + \sqrt{n\kappa_{\max}}) \log \frac{1}{\epsilon})$.
- Lower complexity bound of $\mathcal{O}((n + \sqrt{n\kappa_{\max}}) \log \frac{1}{\epsilon})$ is proved for strongly-convex and smooth finite-sum problems.

SPIDER/SARAH/STORM (VR for non-convex f)

- If objective satisfies *average-smoothness* $\mathbb{E}_i \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|^2 \leq L^2 \|\mathbf{x} - \mathbf{y}\|^2$, then complexity can be reduced to $\mathcal{O}(\min\{\sqrt{n}/\epsilon^2, \epsilon^{-3}\})$ instead of $\mathcal{O}(n/\epsilon^2)$.
- Algorithm
 - Input T iteration, Q epoch length, D batch size 1, D batch size 2, x_0 , *alpha*, η , $\omega(x)$
 - For $t \in [0 : T - 1]$ do
 - If $t \equiv 0 \pmod{Q}$ then
 - compute $\mathbf{g}_t = \frac{1}{D} \sum_{i=1}^D \nabla f(\mathbf{x}_t; \xi_t^i)$
 - else
 - compute $\mathbf{g}_t = (1 - \eta) \left(\mathbf{g}_{t-1} - \frac{1}{S} \sum_{i=1}^S \nabla f(\mathbf{x}_{t-1}; \xi_t^i) \right) + \frac{1}{S} \sum_{i=1}^S \nabla f(\mathbf{x}_t; \xi_t^i)$.
 - End if
 - $\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x} \in X} \{ \mathbf{g}_t^\top \mathbf{x} + \frac{1}{\alpha} V_\omega(\mathbf{x}, \mathbf{x}_t) \}$
 - End for
 - Output \mathbf{x}_τ with τ randomly chosen from $[0 : T - 1]$
- Complexity Total time for gradient calculation is $\mathcal{O}(T(2S + D/Q))$.
- Difference among each algorithms
 - STORM usually is the best.

Parameters	SPIDER	SARAH	STORM	New 1	New 2
T	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-5/2})$	$\mathcal{O}(\epsilon^{-3})$
T/Q	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(\epsilon^{-1})$	1	$\mathcal{O}(\epsilon^{-1})$	1
D	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-2})$	$\mathcal{O}(\epsilon^{-1})$
S	$\mathcal{O}(\epsilon^{-1})$	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon^{-1/2})$	$\mathcal{O}(1)$
η_t	0	0	$\mathcal{O}(t^{-2/3})$	0	$\mathcal{O}(\epsilon^2)$
α_t	$\mathcal{O}(1)$	$\mathcal{O}(\epsilon)$	$\mathcal{O}(t^{-1/3})$	$\mathcal{O}(\epsilon^{1/2})$	$\mathcal{O}(\epsilon)$
Complexity	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$