Chapter 12 Stochastic Optimization

- Formulation $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := \mathbb{E}_{\boldsymbol{\xi}}[f(\mathbf{x},\boldsymbol{\xi})]$ or special case of $\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$.
  - $\boldsymbol{\xi}\in\Xi\subset\mathbb{R}^m$ 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
  i. $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \nabla f\left(\mathbf{x}_t, \boldsymbol{\xi}_t\right)$ where $\boldsymbol{\xi}_t \overset{iid}{\sim} P(\boldsymbol{\xi})$.
  ii. $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \nabla f_{i_t}\left(\mathbf{x}_t\right)$ where $i_t \in [n]$ is uniformly sampled.
- We need $\gamma_t \to 0$ to achieve convergence.
- To reduce noise, we can use *mini-batch* $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \cdot \frac{1}{b}\sum_{j\in J,|J|=b} \nabla f\left(\mathbf{x}_t, \boldsymbol{\xi}_j\right)$ or *SGD with iterate averaging* $\overline{\mathbf{x}}_t = \frac{1}{t}\sum_{\tau=1}^t \mathbf{x}_\tau$.

Convergence for convex functions

- SGD is a special case of *stochastic mirror descent*, $\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in X}\left\{V_\omega\left(\mathbf{x},\mathbf{x}_t\right) + \left\langle \gamma_t G\left(\mathbf{x}_t,\boldsymbol{\xi}_t\right), \mathbf{x}\right\rangle\right\}$.
  - $\mathbb{E}[G(\mathbf{x},\boldsymbol{\xi})] \in \partial F(\mathbf{x})$ and we assume $\mathbb{E}\left[\|G(\mathbf{x},\boldsymbol{\xi})\|_*^2\right] \le M^2$.
- Theorem 12.4 F convex, then SMD gives $\mathbb{E}\left[F\left(\hat{\mathbf{x}}_T\right) - F\left(\mathbf{x}_*\right)\right] \le \frac{R^2 + \frac{M^2}{2}\sum_{t=1}^T \gamma_t^2}{\sum_{t=1}^T \gamma_t}$, where $R^2 = \max_{\mathbf{x}\in X} V_\omega\left(\mathbf{x},\mathbf{x}_1\right)$ 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 $\gamma_t\left(f\left(\mathbf{x}_t,\boldsymbol{\xi}_t\right) - f(\mathbf{x}^*,\boldsymbol{\xi}_t)\right) \le V_\omega\left(\mathbf{x}^*,\mathbf{x}_t\right) - V_\omega\left(\mathbf{x}^*,\mathbf{x}_{t+1}\right) + \frac{\gamma_t^2}{2}\|G\left(\mathbf{x}_t,\boldsymbol{\xi}_t\right)\|_{a*}^2$
      - 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. $\boldsymbol{\xi}_t$ and condition on $\mathbf{x}_t$ we get ($\mathbf{x}_t$ is still stochastic.)
      - $\gamma_t\left(F\left(\mathbf{x}_t\right) - F(\mathbf{x}^*)\right) \le V_\omega\left(\mathbf{x}^*,\mathbf{x}_t\right) - \mathbb{E}\left[V_\omega\left(\mathbf{x}^*,\mathbf{x}_{t+1}\right) \mid \mathbf{x}_t\right] + \frac{\gamma_t^2}{2}M^2$
      - Or equivalently under 2-norm we can get with only convexity of F,
        - $2\gamma_t\left(F\left(\mathbf{x}_t\right) - F(\mathbf{x}^*)\right) \le \|\mathbf{x}^* - \mathbf{x}_t\|_2^2 - \mathbb{E}\left[\|\mathbf{x}^* - \mathbf{x}_{t+1}\|^2 \mid \mathbf{x}_t\right] + \gamma_t^2\|G\left(\mathbf{x}_t,\boldsymbol{\xi}_t\right)\|_2^2$
    - Further taking expectation w.r.t $\mathbf{x}_t$ and condition on $\mathbf{x}_{t-1}$ and we get
      - $\gamma_t\mathbb{E}\left[F\left(\mathbf{x}_t\right) - F(\mathbf{x}^*) \mid \mathbf{x}_{t-1}\right] \le \mathbb{E}\left[V_\omega\left(\mathbf{x}^*,\mathbf{x}_t\right) - V_\omega\left(\mathbf{x}^*,\mathbf{x}_{t+1}\right) \mid \mathbf{x}_{t-1}\right] + \frac{\gamma_t^2}{2}M^2$
      - add w.r.t. $t-1$ case and we get
        $\gamma_{t-1}\left(F\left(\mathbf{x}_{t-1}\right) - F(\mathbf{x}^*)\right) + \gamma_t\mathbb{E}\left[F\left(\mathbf{x}_t\right) - F(\mathbf{x}^*) \mid \mathbf{x}_{t-1}\right] \le V_\omega\left(\mathbf{x}^*,\mathbf{x}_{t-1}\right) - \mathbb{E}\left[V_\omega\left(\mathbf{x}^*,\mathbf{x}_{t+1}\right) \mid \mathbf{x}_t\right] + (\gamma_{t-1}^2 + \gamma_t^2)M^2/2$
    - do this expectation and summation through $t\in[1:T]$ and we get
      - $\sum_{t=1}^T \gamma_t\mathbb{E}\left[F\left(\mathbf{x}_t\right) - F(\mathbf{x}^*) \mid \mathbf{x}_1\right] \le V_\omega\left(\mathbf{x}^*,\mathbf{x}_1\right) - \mathbb{E}\left[V_\omega\left(\mathbf{x}^*,\mathbf{x}_{T+1}\right) \mid \mathbf{x}_t\right] + \sum_{t=1}^T \gamma_t^2 M^2/2$
      - then the conclusion is straightforward.
  - If we set $\gamma_t \equiv \frac{R}{M\sqrt{T}}$, then $\mathbb{E}\left[F\left(\hat{\mathbf{x}}_T\right) - F\left(\mathbf{x}^*\right)\right] = O\left(\frac{BR}{\sqrt{T}}\right)$, this means $O\left(1/\epsilon^2\right)$ of sample complexity.

Convergence for strongly convex functions

- Descent Lemma assume f is μ-strongly convex and $\mathbb{E}\left[\|\nabla f(\mathbf{x},\boldsymbol{\xi})\|_2^2\right] \le B^2$, then
  $\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right] \le (1-2\mu\gamma_t)\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2\right] + \gamma_t^2 B^2$
  - Proof
    - $\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \le \|\mathbf{x}_t - \gamma_t\nabla f\left(\mathbf{x}_t,\boldsymbol{\xi}_t\right) - \mathbf{x}_*\|_2^2 = \|\mathbf{x}_t - \mathbf{x}_*\|_2^2 - 2\gamma_t\left\langle \nabla f\left(\mathbf{x}_t,\boldsymbol{\xi}_t\right), \mathbf{x}_t - \mathbf{x}_*\right\rangle + \gamma_t^2\|\nabla f\left(\mathbf{x}_t,\boldsymbol{\xi}_t\right)\|_2^2$
    - Taking expectation over $\boldsymbol{\xi}_t$ and we get
      - $\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \mid \mathbf{x}_t\right] \le \|\mathbf{x}_t - \mathbf{x}_*\|_2^2 - 2\gamma_t\left\langle \nabla F\left(\mathbf{x}_t\right), \mathbf{x}_t - \mathbf{x}_*\right\rangle + \gamma_t^2 B^2$
    - Strong convexity of F means coercivity, $\left\langle \nabla F\left(\mathbf{x}_t\right), \mathbf{x}_t - \mathbf{x}_*\right\rangle = \left\langle \nabla F\left(\mathbf{x}_t\right) - F\left(\mathbf{x}^\star\right), \mathbf{x}_t - \mathbf{x}_*\right\rangle \ge \mu\|\mathbf{x}_t - \mathbf{x}_*\|_2^2$
      - plug this in and we get the result.
- Theorem 12.3 Assume $F(x)$ is μ-strongly convex, and $\exists B > 0$, s.t. $\forall \mathbf{x}\in X, \mathbb{E}\left[\|\nabla f(\mathbf{x},\boldsymbol{\xi})\|_2^2\right] \le B^2$, then SGD with $\gamma_t = \gamma/t$ at iteration t where $\gamma > 1/2\mu$ satisfies $\mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_*\|_2^2\right] \le \frac{C(\gamma)}{t}$ 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 $\varepsilon_t := \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 \mid \mathbf{x}_1\right]$, we have
      - $\varepsilon_{t+1} \le (1 - 2\mu\gamma/t)\varepsilon_t + \gamma^2 B^2 t^2$.
    - We use induction and assume $\exists C(\gamma)$ s.t. $\forall \tau \le t$, $\varepsilon_\tau \le C(\gamma)/\tau$.
    - To make $\varepsilon_{t+1} \le C(\gamma)/(t+1)$, we need (denote $b := 2\mu\gamma > 1$ and $a := B^2\gamma^2$)
      - $\epsilon_{t+1} \le \frac{C(\gamma)}{t}\left(1 - \frac{b}{t}\right) + \frac{a}{t^2} = \frac{C(\gamma)}{t+1}\left(1 + \frac{1}{t}\right)\left(1 - \frac{b}{t}\right) + \frac{a}{t^2} = \frac{C(\gamma)}{t+1} - \frac{C(\gamma)}{t(t+1)}\left(b - 1 + \frac{b}{t}\right) + \frac{a}{t^2}$
      - 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 indunction succeed, we also need $C(\gamma) \geq \varepsilon_1 = \|\mathbf{x}_1 - \mathbf{x}_*\|_2^2$.
    - sample compexity is $O(1/\epsilon)$.
    - If F also L-smooth, we have $\mathbb{E}\left[F\left(\mathbf{x}_t\right) - F\left(\mathbf{x}^*\right)\right] = O\left(\frac{L \cdot C(\gamma)}{t}\right)$.
- Example SGD over $\min_x F(x) := \frac{1}{2}\mathbb{E}_{\xi\sim N(0,1)}\left[(x-\xi)^2\right]$ with $\gamma_t = 1/t$ gives $\mathbb{E}\left[|x_{t+1} - x^*|^2\right] = \frac{1}{t}$, since $tx_{t+1} = (t-1)x_t + \xi_t$.
- Lemma (Boundedness of Stochastic Gradients, Ex5.1 5.2) $F(x) = \mathbb{E}[f(x,\xi)]$ and f is convex and L-smooth, define $\mathbf{x}^* = \mathbf{argmin}_{\mathbf{x}} F(\mathbf{x})$, then
    - $\mathbb{E}\left[\|\nabla f(x,\xi) - \nabla f\left(x^*,\xi\right)\|_2^2\right] \leq 2L\left[F(x) - F\left(x^*\right)\right]$ and $\mathbb{E}\left[\|\nabla f(x,\xi)\|_2^2\right] \leq 4L\left[F(x) - F\left(x^*\right)\right] + 2\mathbb{E}\left[\|\nabla f\left(x^*,\xi\right)\|_2^2\right]$
    - Proof
      - first
        - Define $f_y(x) := f(x,\xi) - \nabla f(y,\xi)^\top(x-y)$, then $x = y$ is the minima for f_y,
        - by Lemma B in chap3, the quadratic bound for smooth function
          - $f(\mathbf{x}^\star,\xi) - \nabla f(y,\xi)^\top(\mathbf{x}^\star - y) - f(y,\xi) = f_y(\mathbf{x}^\star) - f_y(y) \geq \frac{1}{2L}\|\nabla f_y(\mathbf{x}^\star)\|_2^2 = \frac{1}{2L}\|\nabla f(\mathbf{x}^\star) - \nabla f(\mathbf{x}_y)\|_2^2$.
        - Taking expectation over ξ and $\nabla F(\mathbf{x}^\star) = 0$, we get
          - $F(\mathbf{x}^\star) - F(y) \geq \frac{1}{2L}\mathbb{E}\|\nabla f(\mathbf{x}^\star) - \nabla f(\mathbf{x}_y)\|_2^2$
      - Second
        - by $\|\mathbf{a}+\mathbf{b}\|_2^2 \leq 2\|\mathbf{a}\|_2^2 + 2\|\mathbf{b}\|_2^2$
          - so $\|\nabla f(x,\xi)\|_2^2 \leq 2\|\nabla f(x,\xi) - \nabla f\left(x^*,\xi\right)\|_2^2 + 2\|\nabla f\left(x^*,\xi\right)\|_2^2$.
- Lemma of my own (might be useful in exam) f convex and L-smooth, then $\forall r, f(x) - f(y) \geq \nabla f_y^\top(x-y) - (\nabla f_x - \nabla f_y)^\top r - L\|r\|_2^2/2$
    - Proof
      - By smoothness, $f(x+r) \leq f(x) + \nabla f_x^\top r + L\|r\|_2^2/2$
      - By convexity $f(x+r) \geq f(y) + \nabla f_y^\top(x+r-y)$
      - 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 $\mathbb{E}\left[\|\nabla f(\mathbf{x},\boldsymbol{\xi})\|_2^2\right] \leq \sigma^2 + c\|\nabla F(\mathbf{x})\|_2^2$, then SGD with constant step size $\gamma_t \equiv \gamma \leq \frac{1}{Lc}$ gives $\mathbb{E}\left[F\left(\mathbf{x}_t\right) - F\left(\mathbf{x}_*\right)\right] \leq \frac{\gamma L \sigma^2}{2\mu} + (1-\gamma\mu)^{t-1}\left[F\left(\mathbf{x}_1\right) - F\left(\mathbf{x}_*\right)\right]$.
    - 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 $\mathbb{E}[F(\mathbf{x}_{t+1})|\mathbf{x}_t] - F(\mathbf{x}_t) \leq -\gamma\|\nabla F(\mathbf{x}_t)\|^2 + \frac{L\gamma^2}{2}\left(\sigma^2 + c\|\nabla F(\mathbf{x}_t)\|^2\right)$
      - By strong convexity we have $\|\nabla F(\mathbf{x}_t) - \nabla F(\mathbf{x}^\star)\| \geq 2\mu(F(\mathbf{x}_t) - F(\mathbf{x}^\star))$, take half of this term into above and we get
      - $\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^\star)|\mathbf{x}_t] \leq (1-\gamma\mu)\left(F(\mathbf{x}_t) - F(\mathbf{x}^\star)\right) + \frac{L\gamma^2\sigma^2}{2} - (1 - L\gamma c)\frac{\gamma}{2}\|\nabla F(\mathbf{x}_t)\|^2$
      - Since $\gamma \geq 1/cL$, we get $\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^\star)] \leq (1-\gamma\mu)\mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^\star)] + \frac{L\gamma^2\sigma^2}{2}$
        - $(1-\gamma\mu)^{-T+1}\mathbb{E}[F(\mathbf{x}_{t+1}) - F(\mathbf{x}^\star)] \leq \mathbb{E}[F(\mathbf{x}_t) - F(\mathbf{x}^\star)] + \sum_{t=1}^{T-1}(1-\gamma\mu)^{-t}\frac{L\gamma^2\sigma^2}{2}$, else is easy.
    - If we start from $\|\mathbf{x}_t - \mathbf{x}^\star\|_2^2$ 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 $\mathbb{E}[\|\nabla f(\mathbf{x},\boldsymbol{\xi}) - \nabla F(\mathbf{x})\|_2^2] \leq \sigma^2$, then the condition natually holds with $c = 1$.
    - When $\mathbb{E}\left[\|\nabla f(\mathbf{x},\boldsymbol{\xi})\|_2^2\right] \leq c\|\nabla F(\mathbf{x})\|_2^2$, 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 *imterpolation*, $\nabla f_i\left(\mathbf{x}^*\right) = 0, \forall i$.
      - 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 $f(\mathbf{x},\xi) - f(\mathbf{x}^\star,\xi) \geq \frac{1}{2L}\|\nabla f(\mathbf{x},\xi)\|_2^2 + \mathbf{Linear\ Term}$, taking expectation over ξ
        - $\mathbb{E}\|\nabla f(\mathbf{x},\xi)\|_2^2 \leq 2L(F(\mathbf{x}) - F(\mathbf{x}^\star))$, then by PL of F, $F(\mathbf{x}) - F(\mathbf{x}^\star) \leq \frac{1}{2\mu}\|\nabla F(\mathbf{x})\|_2^2$ we get strong growth condition with $c = L/\mu$.
- Definition (Weak Growth) $F = \mathbb{E}f$ is L-smooth and have a minima $\mathbf{x}^\star$, stochastic gradient satisfies the weak growth condition with

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

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

Convergence for nonconvex functions (handout 11)

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

Lower Bound

- Sample complexity of $O\left(1/\epsilon^2\right)$ and $O(1/\epsilon)$ for convex / strongly convex cannot be improved.
- Definition (Stochastic Oracle) Given input $\mathbf{x}$, stochastic oracle returns $G(\mathbf{x}, \boldsymbol{\xi})$ such that $\mathbb{E}[G(\mathbf{x}, \boldsymbol{\xi})] \in \partial F(\mathbf{x})$ and $\mathbb{E}\left[\|G(\mathbf{x}, \boldsymbol{\xi})\|_p^2\right] \leq M^2$, $M > 0$ 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 $1 \leq p \leq 2$,
     - $\mathbb{E}\left[f\left(\mathbf{x}_T\right) - f\left(\mathbf{x}^*\right)\right] \geq \mathbf{min}\left\{c_0 Mr\sqrt{\frac{d}{T}}, \frac{Mr}{144}\right\}$
  ii. $\exists c_1, c_2 > 0$, μ-strongly convex function f, for any algorithm making T stochastic oracles with $p = 1$,
     - $\mathbb{E}\left[f\left(\mathbf{x}_T\right) - f\left(\mathbf{x}^*\right)\right] \geq \mathbf{min}\left\{c_1\frac{M^2}{\mu^2 T}, c_2 Mr\sqrt{\frac{d}{T}}, \frac{M^2}{1152\mu^2 d}, \frac{Mr}{144}\right\}$.

Adaptive Stochastic Gradient Methods

- Generic Adaptive Scheme
  - $\mathbf{g}_t = \nabla f\left(\mathbf{x}_t, \boldsymbol{\xi}_t\right)$
  - $\mathbf{m}_t = \phi_t\left(\mathbf{g}_1, \ldots, \mathbf{g}_t\right)$
  - $V_t = \psi_t\left(\mathbf{g}_1, \ldots, \mathbf{g}_t\right)$
  - $\hat{\mathbf{x}}_t = \mathbf{x}_t - \alpha_t V_t^{-1/2}\mathbf{m}_t$
  - $\mathbf{x}_{t+1} = \underset{\mathbf{x} \in X}{\mathbf{argmin}}\left\{\left(\mathbf{x} - \hat{\mathbf{x}}_t\right)^T V_t^{1/2}\left(\mathbf{x} - \hat{\mathbf{x}}_t\right)\right\}$
- SGD $\phi_t\left(\mathbf{g}_1, \ldots, \mathbf{g}_t\right) = \mathbf{g}_t, \quad \psi_t\left(\mathbf{g}_1, \ldots, \mathbf{g}_t\right) = \mathbb{I}$
- AdaGrad $\phi_t\left(\mathbf{g}_1, \ldots, \mathbf{g}_t\right) = \mathbf{g}_t, \quad \psi_t\left(\mathbf{g}_1, \ldots, \mathbf{g}_t\right) = \mathbf{diag}\left(\sum_{\tau=1}^t \mathbf{g}_\tau^2\right)/t$
  - can be viewed as Mirror descent with Bregman Divergence of $\omega_t(\mathbf{x}) = \frac{1}{2}\mathbf{x}_t H_t \mathbf{x}$, where $H_t = \epsilon\mathbf{I} + \left[\sum_{t=1}^t \mathbf{g}_t\mathbf{g}_t^T\right]^{\frac{1}{2}}$.
- RMSProp $\phi_t\left(\mathbf{g}_1, \ldots, \mathbf{g}_t\right) = \mathbf{g}_t, \quad \psi_t\left(\mathbf{g}_1, \ldots, \mathbf{g}_t\right) = (1 - \beta_2)\,\mathbf{diag}\left(\sum^t \beta_2^{t-\tau}\mathbf{g}_\tau^2\right)$
  - momentum on gradient norm term to slow down the decay of learning rate.
- Adam $\phi_t\left(\mathbf{g}_1, \ldots, \mathbf{g}_t\right) = (1 - \beta_1)\sum_{\tau=1}^t \beta_1^{t-\tau}\mathbf{g}_\tau, \quad \psi_t\left(\mathbf{g}_1, \ldots, \mathbf{g}_t\right) = (1 - \beta_2)\,\mathbf{diag}\left(\sum^t \beta_2^{t-\tau}\mathbf{g}_\tau^2\right)$,
  - equivalently $\mathbf{m}_t = \beta_1\mathbf{m}_{t-1} + (1 - \beta_1)\mathbf{g}_t, V_t = \beta_2 V_{t-1} + (1 - \beta_2)\,\mathbf{diag}\left(\mathbf{g}_t^2\right)$.
  - 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 $\mathbb{E}\left[\Delta_t\right] \leq 0$.

- A fix: *AMSGrad*, can prove convergence for many convex case.
    - changes: $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2, \hat{v}_t = \mathbf{max}\left(\hat{v}_{t-1}, v_t\right)$ and $\hat{V}_t = \mathbf{diag}\left(\hat{v}_t\right)$.
    - Idea: if $g_t \ll m_t$ then $v_t < v_{t+1}$, this might increase step size.

Variance reduce methods

- Idea $\mathbb{E}\left[F\left(\mathbf{x}_t\right) - F\left(\mathbf{x}^*\right)\right] \leq \frac{\gamma L \sigma^2}{2\mu} + (1-\mu\gamma)^{t-1}\left(F\left(\mathbf{x}_1\right) - F\left(\mathbf{x}^*\right)\right)$, if σ^2 can be reduced -> better upper bound.
- Mini-Batching $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \frac{1}{b} \sum_{i=1}^{b} \nabla f\left(\mathbf{x}_t, \boldsymbol{\xi}_{t,i}\right)$, $\sigma^2 \leftarrow \sigma^2 / b$ but at cost of computation. sample complexity remains the same.
- Importance sampling since $\xi \sim P$, we can change to another distri Q by $G\left(\mathbf{x}_t, \boldsymbol{\xi}_t\right) \Longrightarrow G\left(\mathbf{x}_t, \boldsymbol{\eta}_t\right) \frac{P(\boldsymbol{\eta}_t)}{Q(\boldsymbol{\eta}_t)}$, and variance may be smaller.
- Momentum $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \hat{\mathbf{m}}_t$ where $\hat{\mathbf{m}}_t = c \cdot \sum_{\tau=1}^{t} \alpha^{t-\tau} \nabla f_{i_\tau}\left(\mathbf{x}_\tau\right)$
- 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 $\mathbf{Cov}(X, Y) > 0$
- Point Estimator $\hat{\Theta}_\alpha = \alpha(X - Y) + \mathbb{E}[Y]$, $\mathbb{E}\left[\hat{\Theta}_\alpha\right] = \alpha\mathbb{E}[X] + (1-\alpha)\mathbb{E}[Y]$ and $\mathbb{V}\left[\hat{\Theta}_\alpha\right] = \alpha^2(\mathbb{V}[X] + \mathbb{V}[Y] - 2\,\mathbf{Cov}[X, Y])$.
    - $\alpha = 0 \to 1$, highly bias and no variance $\to$ unbiased but larges variance.
    - If $\mathbf{Cov}[X, Y]$ is sufficiently large, then $\mathbf{Var}\left[\hat{\Theta}_\alpha\right] < \mathbf{Var}[X]$
    - *Idea* $\mathbf{g}_t := \alpha\left(\nabla f_{i_t}\left(\mathbf{x}_t\right) - Y\right) + \mathbb{E}[Y]$ s.t. $\mathbb{E}\left[\left\|\mathbf{g}_t - \nabla F\left(\mathbf{x}_t\right)\right\|^2\right] \to 0$, as $t \to \infty$.
- Choice 1 $Y = \nabla f_{i_t}\left(\mathbf{x}^\star\right)$, $\mathbb{E}[Y] = 0$, unrealistic but concptually useful.
- Choice 2 $Y = \nabla f_{i_t}\left(\overline{\mathbf{x}}_{i_t}\right)$, where $\overline{\mathbf{x}}_{i_t}$ is the last point for with we evaluated $\nabla f_i\left(\overline{\mathbf{x}}_i\right)$. $\mathbb{E}[Y] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i\left(\overline{\mathbf{x}}_i\right)$, but requires storage of $\left\{\overline{\mathbf{x}}_i\right\}_{i=1}^n$ or $\left\{\nabla f_i\left(\overline{\mathbf{x}}_i\right)\right\}_{i=1}^n$
- 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) ($\alpha = \frac{1}{n}, Y = \mathbf{v}_{i_t}$)

- $\mathbf{g}_t = \frac{1}{n}\left(\nabla f_{i_t}\left(\mathbf{x}_t\right) - \mathbf{v}_{i_t}\right) + \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}_i$ where $\mathbf{v}_i$ is the past gradient $\mathbf{v}_i^t = \begin{cases} \nabla f_{i_t}\left(\mathbf{x}_t\right), & \text{if } i = i_t \\ \mathbf{v}_i^{t-1}, & \text{if } i \neq i_t \end{cases}$.
- Equivalently $\mathbf{g}_t = \mathbf{g}_{t-1} - \frac{1}{n}\mathbf{v}_{i_t}^{t-1} + \frac{1}{n}\nabla f_{i_t}\left(\mathbf{x}_t\right)$, 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 $\gamma = 1/\left(16 L_{\max}\right)$ where $L_{\max} := \mathbf{max}_{i \in [n]}\left\{L_i\right\}$, SAG satisfies that $\mathbb{E}\left[F\left(\mathbf{x}_t\right) - F\left(\mathbf{x}^*\right)\right] \leq C \cdot \left(1 - \mathbf{min}\left\{\frac{1}{8n}, \frac{\mu}{16 L_{\max}}\right\}\right)^t$.
- 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)

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

- 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 induition is that $\mathbb{E}\left[\|\mathbf{g}_t - \nabla F(\mathbf{x}_t)\|^2\right] \leq \mathbb{E}\left[\|\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}})\|^2\right] \leq L_{\max}^2 \|\mathbf{x}_t - \tilde{\mathbf{x}}\|^2$
  - closer $\tilde{x}$ to x_t, smaller the variance, so update $\tilde{x}$ in outer loop.
- Algorithm (Two-loop structure)
  - For $s = 1, 2, \ldots$ 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 $t = 0, 1, \ldots, m-1$ do *(inner loop)* ($2m$ grad computation)
      - Randomly pick $i_t \in \{1 : n\}$
      - Update $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\left(\nabla f_{i_t}(\mathbf{x}_t) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})\right)$
    - 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) $f_i(x)$ is convex and L-smooth, then $\frac{1}{n}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{x}^\star)\|_2^2 \leq 2L_{\max}\left(F(\mathbf{x}) - F(\mathbf{x}^\star)\right)$
  - Proof is same as Exercise 6.1.
- Lemma A Denote $\mathbf{g}_t := \nabla f_{i_t}\left(\mathbf{x}^t\right) - \nabla f_{i_t}(\tilde{\mathbf{x}}) + \nabla F(\tilde{\mathbf{x}})$ then $\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right] \leq 4L_{\max}\left[F(\mathbf{x}_t) - F(\mathbf{x}^*) + F(\tilde{\mathbf{x}}) - F(\mathbf{x}^*)\right]$.
  - Proof
    - Define $h_i(x) := f_i(x) - \nabla f_i(\tilde{x})^\top x + \nabla F(\tilde{x})^\top x$, one can check that $H(x) := \mathbb{E}[h_i(x)] = F(x)$
    - By the conclusion in Exercise 6.2, we have $\mathbb{E}[\|\nabla h_i(x_t)\|_2^2] \leq 4L_{\max}[F(x_t) - F(x^\star)] + 2\mathbb{E}[\|\nabla h_i(x^\star)\|_2^2]$
    - By definition $\nabla h_i(x^\star) := (\nabla f_i(x^\star) - \nabla f_i(\tilde{x})) + (\nabla F(\tilde{x}) - \nabla F(x^\star))$
      - Therefore $\mathbb{E}[\|\nabla h_i(x^\star)\|_2^2] = \mathbb{E}[\|\nabla f_i(x^\star) - \nabla f_i(\tilde{x})\|_2^2] + \|\nabla F(\tilde{x}) - \nabla F(x^\star)\|_2^2 + 2\mathbb{E}[\nabla f_i(x^\star) - \nabla f_i(\tilde{x})]^\top(\nabla F(\tilde{x}) - \nabla F(x^\star))$
      - RHS $= \mathbb{E}[\|\nabla f_i(x^\star) - \nabla f_i(\tilde{x})\|_2^2] - \|\nabla F(\tilde{x}) - \nabla F(x^\star)\|_2^2 \leq \mathbb{E}[\|\nabla f_i(x^\star) - \nabla f_i(\tilde{x})\|_2^2] \leq 2L_{\max}(F(\tilde{x}) - F(x^\star))$, by Exercise 6.1.
    - Plug this in and we get the result.
- Theorem 12.11 (Johnson & Zhang, 2013, geometric convergence) Assume $f_i(x)$ 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,
  $\rho = \dfrac{1}{\mu\eta(1-2L\eta)m} + \dfrac{2L\eta}{1-2L\eta} < 1,$
  - then we have geometric convergence in expectation $\mathbb{E}\left[F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}_*)\right] \leq \rho^s\left[F\left(\tilde{\mathbf{x}}^0\right) - F(\mathbf{x}_*)\right]$.
  - Proof
    - $\mathbb{E}\left[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2\right] = \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta(\mathbf{x}_t - \mathbf{x}^*)^T\mathbb{E}\left[\mathbf{g}_t\right] + \eta^2\mathbb{E}\left[\|\mathbf{g}_t\|_2^2\right]$
    - By convexity and Lemma A RHS $\leq \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - 2\eta(1-2L\eta)\left(F(\mathbf{x}_t) - F(\mathbf{x}^*)\right) + 4L\eta^2\left[F(\mathbf{x}_0) - F(\mathbf{x}^*)\right]$
    - The above is for a single step of inner loop, we sum over $t \in [0 : m-1]$ and get
    - $\mathbb{E}\left[\|\mathbf{x}_m - \mathbf{x}^*\|_2^2\right] \leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - 2\eta(1-2L\eta)\sum_{t=0}^{m-1}\mathbb{E}\left[F(\mathbf{x}_t) - F(\mathbf{x}^*)\right] + 4mL\eta^2\left[F(\mathbf{x}_0) - F(\mathbf{x}^*)\right]$
    - By convexity, we have $\sum_{t=0}^{m-1}\left(F(\mathbf{x}_t) - F(\mathbf{x}^*)\right) \geq m\left(F\left(\frac{1}{m}\sum_{t=0}^{m-1}\mathbf{x}_t\right) - F(\mathbf{x}^*)\right) = m\left(F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*)\right)$
    - and by the condition of $\eta < 1/2L$, we have
      - $\mathbb{E}\left[\|\mathbf{x}_m - \mathbf{x}^*\|_2^2\right] \leq \|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - 2\eta m(1-2L\eta)\mathbb{E}\left[F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*)\right] + 4mL\eta^2\left[F(\mathbf{x}_0) - F(\mathbf{x}^*)\right]$
    - By definition $\mathbf{x}_0 = \tilde{\mathbf{x}}^{s-1}$, and by μ-strong convexity $\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 \leq \frac{2}{\mu}(F(\mathbf{x}_0) - F(\mathbf{x}^*))$
      - $2\eta m(1-2L\eta)\mathbb{E}\left[F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}^*)\right] + \underbrace{\mathbb{E}\left[\|\mathbf{x}_m - \mathbf{x}^*\|_2^2\right]}_{\text{omitted}} \leq \left(\frac{2}{\mu} + 4mL\eta^2\right)\mathbb{E}\left[F(\tilde{\mathbf{x}}^{s-1}) - F(\mathbf{x}^*)\right]$
    - This means $\mathbb{E}\left[F(\tilde{\mathbf{x}}^s) - F(\mathbf{x}_*)\right] \leq \left[\frac{1}{\mu\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta}\right]\mathbb{E}\left[F\left(\tilde{\mathbf{x}}^{s-1}\right) - F(\mathbf{x}_*)\right]$. QED
- Remark 12.13 Setting $\eta L = \theta$ gives $\rho = \frac{L}{\mu\theta(1-2\theta)m} + \frac{2\theta}{1-2\theta} = O\left(\frac{L}{\mu m} + \textbf{const.}\right)$, 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 $O\left(\left(n+\sqrt{n\kappa_{\max}}\right)\log\frac{1}{\epsilon}\right)$.
- Lower complexity bound of $O\left(\left(n+\sqrt{n\kappa_{\max}}\right)\log\frac{1}{\epsilon}\right)$ 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 $O\left(\min\left\{\sqrt{n}/\epsilon^2,\epsilon^{-3}\right\}\right)$ instead of $O\left(n/\epsilon^2\right)$.
- 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 $t\equiv 0(\mathrm{mod}Q)$ then
      - compute $\mathbf{g}_t=\frac{1}{D}\sum_{i=1}^{D}\nabla f\left(\mathbf{x}_t;\boldsymbol{\xi}_t^i\right)$
    - else
      - compute $\mathbf{g}_t=(1-\eta)\left(\mathbf{g}_{t-1}-\frac{1}{S}\sum_{i=1}^{S}\nabla f\left(\mathbf{x}_{t-1};\boldsymbol{\xi}_t^i\right)\right)+\frac{1}{S}\sum_{i=1}^{S}\nabla f\left(\mathbf{x}_t;\boldsymbol{\xi}_t^i\right)$.
    - End if
    - $\mathbf{x}_{t+1}=\mathbf{argmin}_{\mathbf{x}\in X}\left\{\mathbf{g}_t^\top x+\frac{1}{\alpha}V_\omega\left(\mathbf{x},\mathbf{x}_t\right)\right\}$
  - End for
  - Output $\mathbf{x}_\tau$ with τ randomly chosen from $[0:T-1]$
- *Complexity* Total time for gradient calculation is $\mathcal{O}(T(2S+D/Q))$.
- Difference amoung each algorithms
  - STORM usually is the best.

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

Stochastic Path Integrated Dierential Estimator (SPIDER) Algorithm (Ex 11)

- Assumption
  i. $\mathbf{min}_{x\in\mathbb{R}^d}F(x)=\mathbb{E}[f(x;\xi)]$, F *not* necessarily convex.
  ii. $\mathbb{E}\left[\|\nabla f(x;\xi)-\nabla F(x)\|^2\right]\leq\sigma^2$,
  iii. $\mathbb{E}\left[\|\nabla f(x;\xi)-\nabla f(y;\xi)\|^2\right]\leq\mathbf{L}^2\|x-y\|^2$
    - This implies smoothness of F, since $\mathbb{E}\left[\|\nabla f(x;\xi)-\nabla f(y;\xi)\|^2\right]\geq\|\mathbb{E}\left[\nabla f(x;\xi)-\nabla f(y;\xi)\right]\|^2$
- Define $\nabla_B F(\mathbf{x}):=\frac{1}{|B|}\sum_{b\in B}\nabla f(x;\mathbf{b})$
- Algorithm
  - Input T iterations, L smoothness, σ^2, ϵ and starting point x_0
  - let $S_1\leftarrow 2\sigma^2/\epsilon^2$, $S_2\leftarrow 2\sigma/\epsilon$, $q\leftarrow\sigma/\epsilon$ are all integers.
  - For $t\in[0:T-1]$ do
    - If $t\bmod q\equiv 0$ then
      - Draw S_1 samples to be B_t and let $v_t\leftarrow\nabla F_{B_t}(x_t)$
    - else
      - Draw S_2 samples to be B_t and let $v_t\leftarrow\nabla F_{B_t}(x_t)-\nabla F_{B_t}(x_{t-1})+v_{t-1}$
    - End If
    - $\eta_t\leftarrow\mathbf{min}\{\epsilon/(L\|v_t\|),1/(2L)\}$
    - $x_{t+1}\leftarrow x_t-\eta_t v_t$
  - End For
  - Return a uniformly random iterate form $0:T-1$.

- Lemma B $F(x_{t+1}) \leq F(x_t) - \frac{\epsilon\|v_t\|}{4L} + \frac{\epsilon^2}{2L} + \frac{\eta_t}{2}\|v_t - \nabla F(x_t)\|^2$
  - Proof
    - By smoothness,
      $F(x_{t+1}) - F(x_t) \leq -\eta_t \nabla F(x_t)^\top v_t + \frac{L}{2}\eta_t^2\|v_t\|^2 = \frac{\eta_t}{2}\|v_t - \nabla F(x_t)\|^2 + \frac{\eta_t\|v_t\|^2}{2}(L\eta_t - 1) - \frac{\eta_t}{2}\|\nabla F(x_t)\|^2$
    - If $\eta = \epsilon/(L\|v_t\|) \leq 1/(2L)$, $\frac{\eta_t\|v_t\|^2}{2}(L\eta_t - 1) = \frac{\epsilon^2}{2L} - \frac{\epsilon\|v_t\|}{2L} \leq \frac{\epsilon^2}{2L} - \frac{\epsilon\|v_t\|}{4L}$, disgard term $\|\nabla F(x_t)\|^2$ we get the equation.
    - If $\eta = 1/(2L) \leq \epsilon/(L\|v_t\|)$, $\frac{\eta_t\|v_t\|^2}{2}(L\eta_t - 1) = \frac{\epsilon\|v_t\|}{2L}(\frac{\epsilon}{\|v_t\|} - 1) = -\frac{\|v_t\|^2}{2L}$,
      - and by fact of $\frac{\epsilon}{\|v_t\|} \geq 1/2$, quadratic function $-\frac{\epsilon\|v_t\|}{4L} + \frac{\epsilon^2}{2L} = \frac{\|v_t\|^2}{2L}(\frac{\epsilon}{\|v_t\|} - 1/2)\frac{\epsilon}{\|v_t\|}$ is always positive,
    - disguard both $\|\nabla F(x_t)\|^2$ and $\frac{\eta_t\|v_t\|^2}{2}(L\eta_t - 1)$ and add $\frac{\epsilon\|v_t\|}{4L} + \frac{\epsilon^2}{2L}$ this term, we get the inequality.
- Lemma C define $k = \lfloor t/q \rfloor \cdot q$ as the last step 1 before t, $\mathbb{E}\left[\|v_t - \nabla f(x_t))\|^2 \mid k\right] \leq \epsilon^2$
  - Proof
    - By definition, for $\tau \in [k+1:t]$, $\|v_\tau - \nabla F(x_t)\|^2 = \|\nabla F_{B_\tau}(x_\tau) - \nabla F(x_\tau) + v_{\tau-1} - \nabla F_{B_\tau}(x_{\tau-1})\|^2$
      $= \|(\nabla F_{B_\tau}(x_\tau) - \nabla F_{B_\tau}(x_{\tau-1})) + (\nabla F(x_{\tau-1}) - \nabla F(x_\tau)) + (v_{\tau-1} - \nabla F(x_{\tau-1}))\|^2$
    - Expand this 3-term quadratic form and take expecation over sampling and condition on τ, we have
      $\textbf{RHS} = \mathbb{E}\left[\|\nabla F_{B_\tau}(x_\tau) - \nabla F_{B_\tau}(x_{\tau-1})\|^2\right] + \|\nabla F(x_{\tau-1}) - \nabla F(x_\tau)\|^2 + \|v_{\tau-1} - \nabla F(x_{\tau-1})\|^2$
      $+2\mathbb{E}\left[\langle \nabla F_{B_\tau}(x_\tau) - \nabla F_{B_\tau}(x_{\tau-1}), \nabla F(x_{\tau-1}) - \nabla F(x_\tau)\rangle\right] + 2\langle \mathbb{E}\left[\nabla F_{B_\tau}(x_\tau) - \nabla F_{B_\tau}(x_{\tau-1})\right]. v_{\tau-1} - \nabla F(x_{\tau-1})\rangle$
      $+2\langle \nabla F(x_{\tau-1}) - \nabla F(x_\tau), v_{\tau-1} - \nabla F(x_{\tau-1})\rangle$
    - The fourth term is twice the negative of the second term, the fifth and sixth term cancel out, so
      $\textbf{RHS} = \|v_{\tau-1} - \nabla F(x_{\tau-1})\|^2 + \mathbb{E}\left[\|\nabla F_{B_\tau}(x_\tau) - \nabla F_{B_\tau}(x_{\tau-1})\|^2\right] - \|\nabla F(x_{\tau-1}) - \nabla F(x_\tau)\|^2$
    - By assumption iii, $\mathbb{E}\left[\|\nabla F_{B_\tau}(x_\tau) - \nabla F_{B_\tau}(x_{\tau-1})\|^2\right] \leq \frac{L^2}{|B_\tau|}\|x_\tau - x_{\tau-1}\|^2$ and ommit the third term, we get
      $\|v_\tau - \nabla F(x_t)\|^2 - \|v_{\tau-1} - \nabla F(x_{\tau-1})\|^2 \leq \frac{L^2}{|B_\tau|}\|x_\tau - x_{\tau-1}\|^2 = \frac{L\eta_{\tau-1}^2}{|B_\tau|}\|v_{\tau-1}\|^2 \leq \frac{\epsilon^2}{|B_\tau|}$
    - Sum and take expectation over $\tau \in [k+1:t]$ we get $\mathbb{E}\left[\|v_\tau - \nabla F(x_t)\|^2 | x_k, v_k\right] \leq \|v_k - \nabla F(x_k)\|^2 + \frac{(t-k-1)\epsilon^2}{|B_\tau|}$
    - By definition $(t-k-1) \leq q = \sigma/\epsilon$, and $|B_\tau| = S_2 = 2\sigma/\epsilon$, then the second term $\frac{(t-k-1)\epsilon^2}{|B_\tau|} \leq \frac{\epsilon^2}{2}$
    - Taking expectation over sampling for the first term, by assumption ii, we get
      $\mathbb{E}\left[\|v_k - \nabla F(x_k)\|^2\right] = \mathbb{E}\left[\|\nabla F_{B_k}(x_k) - \nabla F(x_k)\|^2\right] \leq \frac{1}{|B_k|}\sigma^2 = \frac{\sigma^2}{S_1} = \frac{\epsilon^2}{2}$, sum two term together and we get to the proof.
- Lemma D $\mathbb{E}\left[\|\nabla F(x_t)\|\right] \leq \mathbb{E}\left[\|v_t\|\right] + \epsilon$
  - Proof $\mathbb{E}[\|\nabla F(x_t)\| - \|v_t\|]^2 \leq \mathbb{E}[\|\nabla F(x_t) - v_t\|]^2 \leq \mathbb{E}[\|\nabla F(x_t) - v_t\|^2] \leq \epsilon^2$.
- Lemma E If $F(x_0) - F^\star \leq \delta$, then for $T \geq 4L\delta/\epsilon^2 + 1$, the output is a 5ϵ-approximate first order stationary point, $\mathbb{E}[\|\nabla f(\tilde{x})]\| \leq 5\epsilon$.
  - Proof
    - Sum descent lemma A, for $t = [0:T-1]$, we get $0 \leq \delta - \sum_{t=0}^{T-1} \frac{\epsilon(\epsilon - \mathbb{E}[\|\nabla f(x_t)\|])}{4L} + \frac{\epsilon^2 T}{2L} + \sum_{t=0}^{T-1}\frac{\eta_t\epsilon^2}{2}$
    - So $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(x_t)\|] \leq \epsilon + \frac{4L\delta}{T\epsilon} + 2\epsilon + \frac{2L\epsilon}{T}\sum_{t=0}^{T-1}\eta_t$ since $\eta_t \leq 1/2L$, we have $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(x_t)\|] \leq 4\epsilon + \frac{4L\delta}{T\epsilon}$, by fact of $T \geq 4L\delta/\epsilon^2 + 1$, $\frac{4L\delta}{T\epsilon} \leq \epsilon$, then we get the proof.