Outline

– Lect 1: Recap on convexity
– Lect 1: Recap on duality, optimality
– Lect 2: First-order optimization algorithms
– Lect 3: Operator splitting
– Lect 4: Stochastic and incremental methods


**Large-scale ML**

Regularized Empirical Risk Minimization

\[
\min_w \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^T x_i) + \lambda r(w).
\]

This is the \( f(w) + r(w) \) “composite objective” form we saw.
(e.g., regression, logistic regression, lasso, CRFs, etc.)
Regularized Empirical Risk Minimization

\[
\min_w \quad \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^T x_i) + \lambda r(w).
\]

This is the \( f(w) + r(w) \) “composite objective” form we saw. (e.g., regression, logistic regression, lasso, CRFs, etc.)

- training data: \((x_i, y_i) \in \mathbb{R}^d \times \mathcal{Y}\) (i.i.d.)
- large-scale ML: Both \(d\) and \(n\) are large:
  - \(d\): dimension of each input sample
  - \(n\): number of training data points / samples
- Assume training data “sparse”; so total datasize \(\ll dn\).
- Running time \(O(\#\text{nnz})\)
Regularized Risk Minimization

Empirical: \( \hat{F}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^T x_i) + \lambda r(w) \)

Generalization: \( F(w) = \mathbb{E}_{(x,y)}[\ell(y, w^T x)] + \lambda r(w) \)
Empirical: $\hat{F}(w) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, w^T x_i) + \lambda r(w)$

Generalization: $F(w) = \mathbb{E}_{(x,y)}[\ell(y, w^T x)] + \lambda r(w)$

Single pass through data for $F(w)$ by sampling $n$ iid points

Multiple passes if only minimizing empirical cost $\hat{F}(w)$
Stochastic optimization

\[
\min_{x \in \mathcal{X}} F(x) := \mathbb{E}_{\xi}[f(x, \xi)]
\]

\( (f: \text{loss}; \ x: \text{parameters}; \ \xi: \text{data samples}) \)

Setup

1. \( \mathcal{X} \subset \mathbb{R}^d \) compact convex set
**Stochastic optimization**

$$\min_{x \in \mathcal{X}} F(x) := \mathbb{E}_\xi \[ f(x, \xi) \]$$

($f$: loss; $x$: parameters; $\xi$: data samples)

**Setup**

1. $\mathcal{X} \subset \mathbb{R}^d$ compact convex set
2. $\xi$ r.v. with distribution $P$ on $\Omega \subset \mathbb{R}^d$
Stochastic optimization

\[
\min_{x \in \mathcal{X}} F(x) := \mathbb{E}_\xi[f(x, \xi)] \\
(f: \text{loss}; x: \text{parameters}; \xi: \text{data samples})
\]

Setup

1. \(\mathcal{X} \subset \mathbb{R}^d\) compact convex set
2. \(\xi\) r.v. with distribution \(P\) on \(\Omega \subset \mathbb{R}^d\)
3. The expectation

\[
\mathbb{E}_\xi[f(x, \xi)] = \int_{\Omega} f(x, \xi) dP(\xi)
\]

is well-defined and \textit{finite valued} for every \(x \in \mathcal{X}\).
**Stochastic optimization**

\[
\min_{x \in \mathcal{X}} F(x) := \mathbb{E}_\xi[f(x, \xi)] \\
(f: \text{ loss}; \ x: \text{ parameters}; \ \xi: \text{ data samples})
\]

**Setup**

1. \( \mathcal{X} \subset \mathbb{R}^d \) compact convex set
2. \( \xi \) r.v. with distribution \( P \) on \( \Omega \subset \mathbb{R}^d \)
3. The expectation

\[
\mathbb{E}_\xi[f(x, \xi)] = \int_\Omega f(x, \xi) dP(\xi)
\]

is well-defined and **finite valued** for every \( x \in \mathcal{X} \).
4. For every \( \xi \in \Omega \), \( f(\cdot, \xi) \) is convex
**Stochastic optimization**

**Assumption 1:** Possible to generate iid samples $\xi_1, \xi_2, \ldots$

**Assumption 2:** Oracle yields stochastic gradient $g(x, \xi)$, i.e.,

$$G(x) := \mathbb{E}[g(x, \xi)] \quad \text{s.t.} \quad G(x) \in \partial F(x).$$
Assumption 1: Possible to generate iid samples $\xi_1, \xi_2, \ldots$

Assumption 2: Oracle yields stochastic gradient $g(x, \xi)$, i.e.,

$$G(x) := \mathbb{E}[g(x, \xi)] \quad \text{s.t.} \quad G(x) \in \partial F(x).$$

Theorem Let $\xi \in \Omega$; If $f(\cdot, \xi)$ is convex, and $F(\cdot)$ is finite valued in a neighborhood of $x$, then

$$\partial F(x) = \mathbb{E}[\partial_x f(x, \xi)].$$
Stochastic optimization

**Assumption 1:** Possible to generate iid samples $\xi_1, \xi_2, \ldots$

**Assumption 2:** Oracle yields stochastic gradient $g(x, \xi)$, i.e.,

$$G(x) := \mathbb{E}[g(x, \xi)] \quad \text{s.t.} \quad G(x) \in \partial F(x).$$

**Theorem** Let $\xi \in \Omega$; If $f(\cdot, \xi)$ is convex, and $F(\cdot)$ is finite valued in a neighborhood of $x$, then

$$\partial F(x) = \mathbb{E}[\partial_x f(x, \xi)].$$

$\blacktriangleright$ So $g(x, \omega) \in \partial_x f(x, \omega)$ is a stochastic subgradient.
Stochastic optimization methods

♣ Stochastic Approximation (SA) / Stochastic gradient (SGD)
  ► Sample $\xi$ iid

Sample average approximation (SAA)

► Generate $n$ iid samples, $\xi_1, ..., \xi_n$

► Consider empirical objective
  $\hat{F}_n := \frac{1}{n-1} \sum_{i} f(x, \xi_i)$

SAA refers to creation of this sample average problem

► Minimizing $\hat{F}_n$ still needs to be done!
Stochastic optimization methods

♣ Stochastic Approximation (SA) / Stochastic gradient (SGD)
  ➤ Sample $\xi$ iid
  ➤ Generate stochastic subgradient $g(x, \xi)$
Stochastic optimization methods

♣ Stochastic Approximation (SA) / Stochastic gradient (SGD)
  ► Sample $\xi$ iid
  ► Generate stochastic subgradient $g(x, \xi)$
  ► Use that in a subgradient method
Stochastic optimization methods

♣ Stochastic Approximation (SA) / Stochastic gradient (SGD)
  ► Sample $\xi$ iid
  ► Generate stochastic subgradient $g(x, \xi)$
  ► Use that in a subgradient method

♣ Sample average approximation (SAA)

$\xi$ iid
$g(x, \xi)$
Sample average approximation (SAA)
Stochastic optimization methods

♣ Stochastic Approximation (SA) / Stochastic gradient (SGD)
  ► Sample $\xi$ iid
  ► Generate stochastic subgradient $g(x, \xi)$
  ► Use that in a subgradient method

♣ Sample average approximation (SAA)
  ► Generate $n$ iid samples, $\xi_1, \ldots, \xi_n$
Stochastic optimization methods

♣ Stochastic Approximation (SA) / Stochastic gradient (SGD)
  ▶ Sample $\xi$ iid
  ▶ Generate stochastic subgradient $g(x, \xi)$
  ▶ Use that in a subgradient method

♣ Sample average approximation (SAA)
  ▶ Generate $n$ iid samples, $\xi_1, \ldots, \xi_n$
  ▶ Consider **empirical objective** $\hat{F}_n := n^{-1} \sum_i f(x, \xi_i)$
Stochastic optimization methods

♣ Stochastic Approximation (SA) / Stochastic gradient (SGD)
  ▶ Sample $\xi$ iid
  ▶ Generate stochastic subgradient $g(x, \xi)$
  ▶ Use that in a subgradient method

♣ Sample average approximation (SAA)
  ▶ Generate $n$ iid samples, $\xi_1, \ldots, \xi_n$
  ▶ Consider empirical objective $\hat{F}_n := n^{-1} \sum_i f(x, \xi_i)$
  ▶ SAA refers to creation of this sample average problem
  ▶ Minimizing $\hat{F}_n$ still needs to be done!
Stochastic gradient

SA or stochastic (sub)-gradient

- Let $x_0 \in \mathcal{X}$
- For $k \geq 0$
  - Sample $\xi_k$; compute $g(x_k, \xi_k)$ using oracle
  - Update $x_{k+1} = P_\mathcal{X}(x_k - \alpha_k g(x_k, \xi_k))$, where $\alpha_k > 0$
Stochastic gradient

SA or stochastic (sub)-gradient

- Let $x_0 \in X$
- For $k \geq 0$
  - Sample $\xi_k$; compute $g(x_k, \xi_k)$ using oracle
  - Update $x_{k+1} = P_X(x_k - \alpha_k g(x_k, \xi_k))$, where $\alpha_k > 0$

We’ll simply write

$$x_{k+1} = P_X(x_k - \alpha_k g_k)$$
Stochastic gradient

SA or stochastic (sub)-gradient

► Let $x_0 \in \mathcal{X}$
► For $k \geq 0$
  ○ Sample $\xi_k$; compute $g(x_k, \xi_k)$ using oracle
  ○ Update $x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g(x_k, \xi_k))$, where $\alpha_k > 0$

We’ll simply write

$$x_{k+1} = P_{\mathcal{X}}(x_k - \alpha_k g_k)$$

Does this work?
Convergence Analysis

- $x_k$ depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
Convergence Analysis

- $x_k$ depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, $x_k$ does not depend on $\xi_k$
Convergence Analysis

- $x_k$ depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, $x_k$ does not depend on $\xi_k$
- Subgradient method analysis hinges upon: $||x_k - x^*||^2$
Convergence Analysis

- $x_k$ depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, $x_k$ does not depend on $\xi_k$
- Subgradient method analysis hinges upon: $\|x_k - x^*\|^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$
Convergence Analysis

► $x_k$ depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
► Of course, $x_k$ does not depend on $\xi_k$
► Subgradient method analysis hinges upon: $\|x_k - x^*\|^2$
► Stochastic subgradient hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$

Denote: $R_k := \|x_k - x^*\|^2$ and $r_k := \mathbb{E}[R_k] = \mathbb{E}[\|x_k - x^*\|^2]$
Convergence Analysis

- $x_k$ depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, $x_k$ does not depend on $\xi_k$
- Subgradient method analysis hinges upon: $\|x_k - x^*\|^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$

Denote: $R_k := \|x_k - x^*\|^2$ and $r_k := \mathbb{E}[R_k] = \mathbb{E}[\|x_k - x^*\|^2]$

Bounding $R_{k+1}$

$$R_{k+1} = \|x_{k+1} - x^*\|^2 = \|P_x(x_k - \alpha_k g_k) - P_x(x^*)\|^2$$
Convergence Analysis

- \( x_k \) depends on rvs \( \xi_1, \ldots, \xi_{k-1} \), so itself random
- Of course, \( x_k \) does not depend on \( \xi_k \)
- Subgradient method analysis hinges upon: \( \|x_k - x^*\|^2 \)
- Stochastic subgradient hinges upon: \( \mathbb{E}[\|x_k - x^*\|^2] \)

**Denote:** \( R_k := \|x_k - x^*\|^2 \) and \( r_k := \mathbb{E}[R_k] = \mathbb{E}[\|x_k - x^*\|^2] \)

**Bounding** \( R_{k+1} \)

\[
R_{k+1} = \|x_{k+1} - x^*\|^2_2 = \|P_{\mathcal{C}}(x_k - \alpha_k g_k) - P_{\mathcal{C}}(x^*)\|^2_2 \\
\leq \|x_k - x^* - \alpha_k g_k\|^2_2
\]
Convergence Analysis

- $x_k$ depends on rvs $\xi_1, \ldots, \xi_{k-1}$, so itself random
- Of course, $x_k$ does not depend on $\xi_k$
- Subgradient method analysis hinges upon: $\|x_k - x^*\|^2$
- Stochastic subgradient hinges upon: $\mathbb{E}[\|x_k - x^*\|^2]$

Denote: $R_k := \|x_k - x^*\|^2$ and $r_k := \mathbb{E}[R_k] = \mathbb{E}[\|x_k - x^*\|^2]$

Bounding $R_{k+1}$

\[
R_{k+1} = \|x_{k+1} - x^*\|^2 \\
\leq \|x_k - x^* - \alpha_k g_k\|^2 \\
= R_k + \alpha_k^2 \|g_k\|^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle.
\]
Convergence analysis

\[ R_{k+1} \leq R_k + \alpha_k^2 \|g_k\|_2^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \]
Convergence analysis

\[
R_{k+1} \leq R_k + \alpha_k^2 \|g_k\|^2_2 - 2\alpha_k \langle g_k, x_k - x^* \rangle
\]

- **Assume:** \( \|g_k\|_2 \leq M \) on \( \mathcal{X} \)

- **Taking expectation:**

\[
R_{k+1} \leq R_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle g_k, x_k - x^* \rangle].
\]
Convergence analysis

\[ R_{k+1} \leq R_k + \alpha_k^2 \|g_k\|^2_2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \]

- **Assume:** \( \|g_k\|_2 \leq M \) on \( \mathcal{X} \)
- **Taking expectation:**
  \[ r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle g_k, x_k - x^* \rangle] \]
- **We need to now get a handle on the last term**
Convergence analysis

\[ R_{k+1} \leq R_k + \alpha_k^2 \|g_k\|_2^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \]

- **Assume:** \( \|g_k\|_2 \leq M \) on \( \mathcal{X} \)
- **Taking expectation:**
  \[ r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle g_k, x_k - x^* \rangle]. \]
- **We need to now get a handle on the last term**
- **Since** \( x_k \) **is independent of** \( \xi_k \), **we have**

\[ \mathbb{E}[\langle x_k - x^*, g(x_k, \xi_k) \rangle] = \]
Convergence analysis

\[ R_{k+1} \leq R_k + \alpha_k^2 \|g_k\|_2^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \]

- **Assume:** \( \|g_k\|_2 \leq M \) on \( \mathcal{X} \)
- **Taking expectation:**
  \[ r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E} [\langle g_k, x_k - x^* \rangle] . \]

- **We need to now get a handle on the last term**
- **Since** \( x_k \) **is independent of** \( \xi_k \), **we have**

\[
\mathbb{E} \left[ \langle x_k - x^*, g(x_k, \xi_k) \rangle \right] = \mathbb{E} \left\{ \mathbb{E} \left[ \langle x_k - x^*, g(x_k, \xi_k) \rangle \mid \xi[1..(k-1)] \right] \right\} =
\]
Convergence analysis

\[ R_{k+1} \leq R_k + \alpha_k^2 \|g_k\|^2_2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \]

- **Assume:** \( \|g_k\|_2 \leq M \) on \( \mathcal{X} \)
- **Taking expectation:**
  \[ r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle g_k, x_k - x^* \rangle]. \]
- **We need to now get a handle on the last term**
- **Since** \( x_k \) **is independent of** \( \xi_k \), we have

\[
\mathbb{E}[\langle x_k - x^*, g(x_k, \xi_k) \rangle] = \mathbb{E}\left\{ \mathbb{E}[\langle x_k - x^*, g(x_k, \xi_k) \rangle | \xi_{[1..(k-1)]}] \right\} \\
= \mathbb{E}\left\{ \langle x_k - x^*, \mathbb{E}[g(x_k, \xi_k) | \xi_{[1..(k-1)]}] \rangle \right\} \\
= \\
\]
Convergence analysis

\[ R_{k+1} \leq R_k + \alpha_k^2 \|g_k\|_2^2 - 2\alpha_k \langle g_k, x_k - x^* \rangle \]

- **Assume:** \( \|g_k\|_2 \leq M \) on \( \mathcal{X} \)
- **Taking expectation:**
  \[ r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle g_k, x_k - x^* \rangle]. \]
- **We need to now get a handle on the last term**
- **Since** \( x_k \) **is independent of** \( \xi_k \), **we have**

\[
\mathbb{E}[\langle x_k - x^*, g(x_k, \xi_k) \rangle] = \mathbb{E}\left\{ \mathbb{E}[\langle x_k - x^*, g(x_k, \xi_k) \rangle | \xi_{[1..(k-1)]}] \right\}
= \mathbb{E}\left\{ \langle x_k - x^*, \mathbb{E}[g(x_k, \xi_k) | \xi_{[1..(k-1)]}] \rangle \right\}
= \mathbb{E}[\langle x_k - x^*, G_k \rangle], \quad G_k \in \partial F(x_k).
\]
Convergence analysis

It remains to bound: $\mathbb{E}[\langle x_k - x^*, G_k \rangle]$
Convergence analysis

It remains to bound: \( \mathbb{E}[\langle x_k - x^*, G_k \rangle] \)

- Since \( F \) is cvx, \( F(x) \geq F(x_k) + \langle G_k, x - x_k \rangle \) for any \( x \in \chi \).
Convergence analysis

It remains to bound: $\mathbb{E}[\langle x_k - x^*, G_k \rangle]$

- Since $F$ is cvx, $F(x) \geq F(x_k) + \langle G_k, x - x_k \rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular

$$2\alpha_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\alpha_k \mathbb{E}[\langle G_k, x^* - x_k \rangle]$$
Convergence analysis

It remains to bound: \( \mathbb{E}[\langle x_k - x^*, G_k \rangle] \)

- Since \( F \) is cvx, \( F(x) \geq F(x_k) + \langle G_k, x - x_k \rangle \) for any \( x \in \mathcal{X} \).
- Thus, in particular

\[
2\alpha_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\alpha_k \mathbb{E}[\langle G_k, x^* - x_k \rangle]
\]

Plug this bound back into the \( r_{k+1} \) inequality:

\[
r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle G_k, x_k - x^* \rangle]
\]
Convergence analysis

It remains to bound: $\mathbb{E}[\langle x_k - x^*, G_k \rangle]$

- Since $F$ is cvx, $F(x) \geq F(x_k) + \langle G_k, x - x_k \rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular

$$2\alpha_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\alpha_k \mathbb{E}[\langle G_k, x^* - x_k \rangle]$$

Plug this bound back into the $r_{k+1}$ inequality:

$$r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle G_k, x_k - x^* \rangle]$$

$$2\alpha_k \mathbb{E}[\langle G_k, x_k - x^* \rangle] \leq r_k - r_{k+1} + \alpha_k M^2$$
Convergence analysis

It remains to bound: \( \mathbb{E}[\langle x_k - x^*, G_k \rangle] \)

- Since \( F \) is cvx, \( F(x) \geq F(x_k) + \langle G_k, x - x_k \rangle \) for any \( x \in \mathcal{X} \).
- Thus, in particular

\[
2\alpha_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\alpha_k \mathbb{E}[\langle G_k, x^* - x_k \rangle]
\]

Plug this bound back into the \( r_{k+1} \) inequality:

\[
\begin{align*}
    r_{k+1} & \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle G_k, x_k - x^* \rangle] \\
    2\alpha_k \mathbb{E}[\langle G_k, x_k - x^* \rangle] & \leq r_k - r_{k+1} + \alpha_k M^2 \\
    2\alpha_k \mathbb{E}[F(x_k) - F(x^*)] & \leq r_k - r_{k+1} + \alpha_k M^2.
\end{align*}
\]
Convergence analysis

It remains to bound: $\mathbb{E}[\langle x_k - x^*, G_k \rangle]$ 

- Since $F$ is cvx, $F(x) \geq F(x_k) + \langle G_k, x - x_k \rangle$ for any $x \in \mathcal{X}$.
- Thus, in particular

$$2\alpha_k \mathbb{E}[F(x^*) - F(x_k)] \geq 2\alpha_k \mathbb{E}[\langle G_k, x^* - x_k \rangle]$$

Plug this bound back into the $r_{k+1}$ inequality:

$$r_{k+1} \leq r_k + \alpha_k^2 M^2 - 2\alpha_k \mathbb{E}[\langle G_k, x_k - x^* \rangle]$$

$$2\alpha_k \mathbb{E}[\langle G_k, x_k - x^* \rangle] \leq r_k - r_{k+1} + \alpha_k M^2$$

$$2\alpha_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2.$$

We’ve bounded the expected progress; What now?
Convergence analysis

\[ 2\alpha_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2. \]
Convergence analysis

\[ 2\alpha_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2. \]

Sum up over \( i = 1, \ldots, k \), to obtain

\[ \sum_{i=1}^{k} (2\alpha_i \mathbb{E}[F(x_i) - f(x^*)]) \leq r_1 - r_{k+1} + M^2 \sum_i \alpha_i^2 \]
Convergence analysis

\[ 2\alpha_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2. \]

Sum up over \( i = 1, \ldots, k \), to obtain

\[
\sum_{i=1}^{k} (2\alpha_i \mathbb{E}[F(x_i) - f(x^*)]) \leq r_1 - r_{k+1} + M^2 \sum_i \alpha_i^2 \\
\leq r_1 + M^2 \sum_i \alpha_i^2.
\]
Convergence analysis

\[ 2\alpha_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2. \]

Sum up over \( i = 1, \ldots, k \), to obtain

\[ \sum_{i=1}^{k} (2\alpha_i \mathbb{E}[F(x_i) - f(x^*)]) \leq r_1 - r_{k+1} + M^2 \sum_i \alpha_i^2 \]

\[ \leq r_1 + M^2 \sum_i \alpha_i^2. \]

Divide both sides by \( \sum_i \alpha_i \), so
Convergence analysis

\[ 2\alpha_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2. \]

Sum up over \( i = 1, \ldots, k \), to obtain

\[ \sum_{i=1}^{k} (2\alpha_i \mathbb{E}[F(x_i) - f(x^*)]) \leq r_1 - r_{k+1} + M^2 \sum_i \alpha_i^2 \]

\[ \leq r_1 + M^2 \sum_i \alpha_i^2. \]

Divide both sides by \( \sum_i \alpha_i \), so

- Set \( \gamma_i = \frac{\alpha_i}{\sum_i \alpha_i} \).
- Thus, \( \gamma_i \geq 0 \) and \( \sum_i \gamma_i = 1 \).
Convergence analysis

\[ 2\alpha_k \mathbb{E}[F(x_k) - F(x^*)] \leq r_k - r_{k+1} + \alpha_k M^2. \]

Sum up over \( i = 1, \ldots, k \), to obtain

\[
\sum_{i=1}^{k} (2\alpha_i \mathbb{E}[F(x_i) - f(x^*)]) \leq r_1 - r_{k+1} + M^2 \sum_i \alpha_i^2 \\
\leq r_1 + M^2 \sum_i \alpha_i^2.
\]

Divide both sides by \( \sum_i \alpha_i \), so

▶ Set \( \gamma_i = \frac{\alpha_i}{\sum_i \alpha_i} \).

▶ Thus, \( \gamma_i \geq 0 \) and \( \sum_i \gamma_i = 1 \)

\[
\mathbb{E} \left[ \sum_i \gamma_i (F(x_i) - F(x^*)) \right] \leq \frac{r_1 + M^2 \sum_i \alpha_i^2}{2 \sum_i \alpha_i}
\]
But we wish to say something about $x_k$
Convergence analysis

- But we wish to say something about $x_k$
- Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$

Easier to talk about averaged $\bar{x}_k := \sum_i^k \gamma_i x_i$.

$f(\bar{x}_k) \leq \sum_i^k \gamma_i F(x_i)$ due to convexity.

So we finally obtain the inequality:

$$E\left[ F(\bar{x}_k) - F(x^*) \right] \leq r_1 + M_2 \sum_i^k \alpha_i^2 \sum_i^k \alpha_i.$$
But we wish to say something about $x_k$
Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$
Easier to talk about averaged

$$\bar{x}_k := \sum_i^k \gamma_i x_i.$$
Convergence analysis

- But we wish to say something about $x_k$
- Since $\gamma_i \geq 0$ and $\sum_i^k \gamma_i = 1$, and we have $\gamma_i F(x_i)$
- Easier to talk about averaged

$$\bar{x}_k := \sum_i^k \gamma_i x_i.$$  

- $f(\bar{x}_k) \leq \sum_i \gamma_i F(x_i)$ due to convexity
Convergence analysis

► But we wish to say something about \( x_k \)
► Since \( \gamma_i \geq 0 \) and \( \sum_i^k \gamma_i = 1 \), and we have \( \gamma_i F(x_i) \)
► Easier to talk about averaged

\[
\bar{x}_k := \sum_i^k \gamma_i x_i.
\]

► \( f(\bar{x}_k) \leq \sum_i \gamma_i F(x_i) \) due to convexity
► So we finally obtain the inequality

\[
\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{r_1 + M^2 \sum_i \alpha_i^2}{2 \sum_i \alpha_i}.
\]
Let $D_{\mathcal{X}} : = \max_{x \in \mathcal{X}} \| x - x^* \|_2$ (act. only need $\| x_1 - x^* \| \leq D_{\mathcal{X}}$)

Assume $\alpha_i = \alpha$ is a constant. Observe that

$$\mathbb{E} [ F(\bar{x}_k) - F(x^*) ] \leq \frac{D_{\mathcal{X}}^2 + M^2 k \alpha^2}{2k\alpha}$$

Minimize rhs over $\alpha > 0$; thus $\mathbb{E} [ F(\bar{x}_k) - F(x^*) ] \leq \frac{D_{\mathcal{X}} M}{\sqrt{k}}$

If $k$ is not fixed in advance, then choose

$$\alpha_i = \frac{\theta D_{\mathcal{X}}}{M \sqrt{i}}, \quad i = 1, 2, \ldots$$

We showed $O(1/\sqrt{k})$ rate
**Theorem** Let \( f(x, \xi) \) be \( C^1_L \) convex. Let \( e_k := \nabla F(x_k) - g_k \) satisfy \( \mathbb{E}[e_k] = 0 \). Let \( \|x_i - x^*\| \leq D \). Also, let \( \alpha_i = 1/(L + \eta_i) \). Then,

\[
\mathbb{E}\left[\sum_{i=1}^{k} F(x_{i+1}) - F(x^*)\right] \leq \frac{D^2}{2\alpha_k} + \sum_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.
\]
Theorem Let \( f(x, \xi) \) be \( C^1_L \) convex. Let \( e_k := \nabla F(x_k) - g_k \) satisfy \( \mathbb{E}[e_k] = 0 \). Let \( \|x_i - x^*\| \leq D \). Also, let \( \alpha_i = 1/(L + \eta_i) \). Then,

\[
\mathbb{E}\left[\sum_{i=1}^{k} F(x_{i+1}) - F(x^*)\right] \leq \frac{D^2}{2\alpha_k} + \sum_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.
\]

As before, by using \( \tilde{x}_k = \frac{1}{k} \sum_{i=1}^{k} x_{i+1} \) we get

\[
\mathbb{E}[F(\tilde{x}_k) - F(x^*)] \leq \frac{D^2}{2\alpha_k k} + \frac{1}{k} \sum_{i=1}^{k} \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.
\]
Theorem Let $f(x, \xi)$ be $C^1_L$ convex. Let $e_k := \nabla F(x_k) - g_k$ satisfy $\mathbb{E}[e_k] = 0$. Let $\|x_i - x^*\| \leq D$. Also, let $\alpha_i = 1/(L + \eta_i)$. Then,

$$\mathbb{E} \left[ \sum_{i=1}^k F(x_{i+1}) - F(x^*) \right] \leq \frac{D^2}{2\alpha_k} + \sum_{i=1}^k \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$ 

As before, by using $\bar{x}_k = \frac{1}{k} \sum_{i=1}^k x_{i+1}$ we get

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{D^2}{2\alpha_k k} + \frac{1}{k} \sum_{i=1}^k \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$ 

Using $\alpha_i = L + \eta_i$ where $\eta_i \propto 1/\sqrt{i}$ we obtain

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] = O\left(\frac{LD^2}{k}\right) + O\left(\frac{\sigma D}{\sqrt{k}}\right)$$

where $\sigma$ bounds the variance $\mathbb{E}[\|e_i\|^2]$
Stochastic optimization – strongly convex

**Theorem** Suppose $f(x, \xi)$ are convex and $F(x)$ is $\mu$-strongly convex. Let $\bar{x}_k := \sum_{i=0}^{k-1} \theta_i x_i$, where $\theta_i = \frac{2(i+1)}{(k+1)(k+2)}$, we obtain

$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] \leq \frac{2M^2}{\mu(k+1)}.$$  

(*Lacoste-Julien, Schmidt, Bach (2012)*)

With uniform averaging $\bar{x}_k = \frac{1}{k} \sum_i x_i$, we get $O(\log k / k)$. 
### SGD convergence summary

<table>
<thead>
<tr>
<th>Cvx Class</th>
<th>Rate</th>
<th>Iterate</th>
<th>Minimax</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C^0_L$</td>
<td>$1/\sqrt{k}$</td>
<td>$\bar{x}_k$</td>
<td>Yes</td>
</tr>
<tr>
<td>$C^0_L$</td>
<td>$\log k/\sqrt{k}$</td>
<td>$x_k$</td>
<td>No</td>
</tr>
<tr>
<td>$C^1_L$</td>
<td>$1/\sqrt{k}$</td>
<td>$\bar{x}_k$</td>
<td>Yes</td>
</tr>
<tr>
<td>$S^0_L$</td>
<td>$(\log k)/k$</td>
<td>$\bar{x}_k, x_k$</td>
<td>No</td>
</tr>
<tr>
<td>$S^1_L$</td>
<td>$1/k$</td>
<td>$\bar{x}_k, x_k$</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Extensions

- Proximal stochastic gradient

\[ x_{k+1} = \text{prox}_{\alpha_k h}[x_k - \alpha_k g(x_k, \xi_k)] \]

(*Xiao 2010; Hu et al. 2009*)

Accelerated versions also possible

(*Ghadimi, Lan (2013)*)

- Related methods:
  - Regularized dual averaging (*Nesterov, 2009; Xiao 2010*)
  - Stochastic mirror-prox (*Nemirovski et al. 2009*)

- ...
SAA / Batch problem

\[
\min F(x) = \mathbb{E}[f(x, \xi)]
\]

Sample Average Approximation (SAA):

- Collect samples \(\xi_1, \ldots, \xi_n\)
- Empirical objective: \(\hat{F}(x) := \frac{1}{n} \sum_{i=1}^{n} f(x, \xi_i)\)
- aka *Empirical Risk Minimization*
SAA / Batch problem

\[
\min F(x) = \mathbb{E}[f(x, \xi)]
\]

Sample Average Approximation (SAA):

- Collect samples \(\xi_1, \ldots, \xi_n\)
- **Empirical objective:** \(\hat{F}(x) := \frac{1}{n} \sum_{i=1}^{n} f(x, \xi_i)\)
- aka *Empirical Risk Minimization*

**Note:** we often optimize \(\hat{F}\) using stochastic subgradient; but theoretical guarantees are then only on the *empirical* suboptimality \(E[\hat{F}(\bar{x}_k)] \leq \ldots\)
Sample Average Approximation (SAA):

- Collect samples $\xi_1, \ldots, \xi_n$
- **Empirical objective:** $\hat{F}(x) := \frac{1}{n} \sum_{i=1}^{n} f(x, \xi_i)$
- aka *Empirical Risk Minimization*
- **Note:** we often optimize $\hat{F}$ using stochastic subgradient; but theoretical guarantees are then only on the *empirical* suboptimality $E[\hat{F}(\bar{x}_k)] \leq \ldots$
- For guarantees on $F(\bar{x}_k)$ more work (*regularization* + concentration)
Finite-sum problems

\[
\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x).
\]
Finite-sum problems

\[ \min_{x \in \mathbb{R}^d} f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \]

Gradient / subgradient methods

\[
\begin{align*}
x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\
x_{k+1} &= x_k - \alpha_k g(x_k), \quad g \in \partial f(x_k) \\
x_{k+1} &= \text{prox}_{\alpha_k r}(x_k - \alpha_k \nabla f(x_k))
\end{align*}
\]
At iteration $k$, we randomly pick an integer

$$i(k) \in \{1, 2, \ldots, m\}$$

$$x_{k+1} = x_k - \alpha_k \nabla f_{i(k)}(x_k)$$

- The update requires only gradient for $f_{i(k)}$
- Uses unbiased estimate $\mathbb{E}[\nabla f_{i(k)}] = \nabla f$
- One iteration now $n$ times faster using $\nabla f(x)$
- But how many iterations do we need?
<table>
<thead>
<tr>
<th>Method</th>
<th>Assumptions</th>
<th>Full</th>
<th>Stochastic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgradient</td>
<td>convex</td>
<td>$O(1/\sqrt{k})$</td>
<td>$O(1/\sqrt{k})$</td>
</tr>
<tr>
<td>Subgradient</td>
<td>strongly cvx</td>
<td>$O(1/k)$</td>
<td>$O(1/k)$</td>
</tr>
</tbody>
</table>

So using stochastic subgradient, solve $n$ times faster.
Stochastic gradient

<table>
<thead>
<tr>
<th>Method</th>
<th>Assumptions</th>
<th>Full</th>
<th>Stochastic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgradient</td>
<td>convex</td>
<td>$O(1/\sqrt{k})$</td>
<td>$O(1/\sqrt{k})$</td>
</tr>
<tr>
<td>Subgradient</td>
<td>strongly cvx</td>
<td>$O(1/k)$</td>
<td>$O(1/k)$</td>
</tr>
</tbody>
</table>

So using stochastic subgradient, solve $n$ times faster.

<table>
<thead>
<tr>
<th>Method</th>
<th>Assumptions</th>
<th>Full</th>
<th>Stochastic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient</td>
<td>convex</td>
<td>$O(1/k)$</td>
<td>$O(1/\sqrt{k})$</td>
</tr>
<tr>
<td>Gradient</td>
<td>strongly cvx</td>
<td>$O((1 - \mu/L)^k)$</td>
<td>$O(1/k)$</td>
</tr>
</tbody>
</table>

– For smooth problems, stochastic gradient needs more iterations
– Widely used in ML, rapid initial convergence
– Several speedup techniques studied, but worst case remains same
Hybrid methods

- Hybrid of stochastic gradient with full gradient.

**Stochastic Average Gradient (SAG)** (Le Roux, Schmidt, Bach 2012)

- Store the gradients of $\nabla f_i$ for $i = 1, \ldots, n$
- Select uniformly at random $i(k) \in \{1, \ldots, n\}$
- Perform the update

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} y_i^k \quad y_i^k = \begin{cases} \nabla f_i(x_k) & \text{if } i = i(k) \\ y_i^{k-1} & \text{otherwise.} \end{cases}$$
Hybrid methods

- Hybrid of stochastic gradient with full gradient.

**Stochastic Average Gradient (SAG)** (Le Roux, Schmidt, Bach 2012)

- store the gradients of $\nabla f_i$ for $i = 1, \ldots, n$
- Select uniformly at random $i(k) \in \{1, \ldots, n\}$
- Perform the update

$$x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} y_i^k \quad y_i^k = \begin{cases} \nabla f_i(x_k) & \text{if } i = i(k) \\ y_i^{k-1} & \text{otherwise.} \end{cases}$$


- Storage overhead; acceptable in some ML settings:
  - $f_i(x) = \ell(l_i, x^T \Phi(a_i))$, $\nabla f_i(x) = \nabla \ell(l_i, x^T \Phi(a_i))\Phi(a_i)$
  - Store only $n$ scalars (since depends only on $x^T a_i$)
### Method Assumptions Rate

<table>
<thead>
<tr>
<th>Method</th>
<th>Assumptions</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient</td>
<td>convex</td>
<td>$O(1/k)$</td>
</tr>
<tr>
<td>Gradient</td>
<td>strongly cvx</td>
<td>$O((1 - \frac{\mu}{L})^k)$</td>
</tr>
<tr>
<td>Stochastic</td>
<td>strongly cvx</td>
<td>$O(1/k)$</td>
</tr>
<tr>
<td>SAG</td>
<td>strongly convex</td>
<td>$O((1 - \min{\frac{\mu}{n}, \frac{1}{8n}})^k)$</td>
</tr>
</tbody>
</table>

This speedup also observed in practice

**Complicated convergence analysis**

**Similar rates for many other methods**

- stochastic dual coordinate (SDCA); [Shalev-Shwartz, Zhang, 2013]
- stochastic variance reduced gradient (SVRG); [Johnson, Zhang, 2013]
- proximal SVRG [Xiao, Zhang, 2014]
- hybrid of SAG and SVRG, SAGA (also proximal); [Defazio et al, 2014]
- accelerated versions [Lin, Mairal, Harchoui; 2015]
- asynchronous hybrid SVRG [Reddi et al. 2015]
- incremental Newton method, S2SGD and MS2GD, ...