Optimization for Machine Learning

(Large-scale methods - A)

SUVRIT SRA
LIDS, Massachusetts Institute of Technology

PKU Summer School on Data Science (July 2017)

Large-scale ML

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.)

Large-scale ML

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

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

Regularized Risk Minimization

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

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 $\widehat{F}(w)$

Stochastic optimization

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

(f: loss; x: parameters; ξ: 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; ξ: data samples)

Setup

1. $\mathcal{X} \subset \mathbb{R}^d$ compact convex set

2. ξ 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: loss; x: parameters; ξ: data samples)

Setup

1. $\mathcal{X} \subset \mathbb{R}^d$ compact convex set
2. ξ 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 $x \in \mathcal{X}$.

Stochastic optimization

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

Setup

1. $\mathcal{X} \subset \mathbb{R}^d$ compact convex set
2. ξ 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.,

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

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

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

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

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

▶ So $g(x, \omega) \in \partial_x f(x, \omega)$ is a stochastic subgradient.

Stochastic optimization methods

- ♣ Stochastic Approximation (SA) / Stochastic gradient (SGD)
  - ▶ Sample ξ iid

Stochastic optimization methods

♣ Stochastic Approximation (SA) / Stochastic gradient (SGD)

- Sample ξ iid
- Generate stochastic subgradient $g(x, \xi)$

Stochastic optimization methods

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

Stochastic optimization methods

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

Stochastic optimization methods

- ♣ Stochastic Approximation (SA) / Stochastic gradient (SGD)
  - ▶ Sample ξ 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 ξ 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 ξ 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 ξ_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 $x_0 \in \mathcal{X}$
- For $k \geq 0$
  - Sample ξ_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

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

Stochastic gradient

SA or stochastic (sub)-gradient

- Let $x_0 \in \mathcal{X}$
- For $k \geq 0$
  - Sample ξ_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

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

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 ξ_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 ξ_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 ξ_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 ξ_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 ξ_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}

$$R_{k+1} = \|x_{k+1} - x^*\|_2^2 = \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2$$

Convergence Analysis

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

$$\begin{aligned} R_{k+1} &= \|x_{k+1} - x^*\|_2^2 = \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2 \\ &\leq \|x_k - x^* - \alpha_k g_k\|_2^2 \end{aligned}$$

Convergence Analysis

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

$$\begin{aligned} R_{k+1} &= \|x_{k+1} - x^*\|_2^2 = \|P_{\mathcal{X}}(x_k - \alpha_k g_k) - P_{\mathcal{X}}(x^*)\|_2^2 \\ &\leq \|x_k - x^* - \alpha_k g_k\|_2^2 \\ &= R_k + \alpha_k^2 \|g_k\|_2^2 - 2\alpha_k \langle g_k,\, x_k - x^* \rangle. \end{aligned}$$

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 ξ_k, we have

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

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 ξ_k, we have

$$\begin{aligned} \mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k)\rangle] &= \mathbb{E}\left\{\mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k)\rangle \mid \xi_{[1..(k-1)]}]\right\} \\ &= \end{aligned}$$

Convergence analysis

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

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

$$\begin{aligned}
\mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k) \rangle] &= \mathbb{E}\left\{ \mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k) \rangle \mid \xi_{[1..(k-1)]}] \right\} \\
&= \mathbb{E}\left\{ \langle x_k - x^*,\, \mathbb{E}[g(x_k, \xi_k) \mid \xi_{[1..(k-1)]}] \rangle \right\} \\
&=
\end{aligned}$$

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 ξ_k, we have

$$\begin{aligned} \mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k) \rangle] &= \mathbb{E}\left\{ \mathbb{E}[\langle x_k - x^*,\, g(x_k, \xi_k) \rangle \mid \xi_{[1..(k-1)]}] \right\} \\ &= \mathbb{E}\left\{ \langle x_k - x^*,\, \mathbb{E}[g(x_k, \xi_k) \mid \xi_{[1..(k-1)]}] \rangle \right\} \\ &= \mathbb{E}[\langle x_k - x^*,\, G_k \rangle], \quad G_k \in \partial F(x_k). \end{aligned}$$

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 $x \in \mathcal{X}$.

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:

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

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:

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

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:

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

We've bounded the expected progress; What now?

Convergence analysis

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

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

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

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

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

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

$$\begin{aligned} \sum\nolimits_{i=1}^{k} (2\alpha_i \mathbb{E}[F(x_i) - f(x^*)]) &\leq r_1 - r_{k+1} + M^2 \sum\nolimits_i \alpha_i^2 \\ &\leq r_1 + M^2 \sum\nolimits_i \alpha_i^2. \end{aligned}$$

Divide both sides by $\sum_i \alpha_i$, so

Convergence analysis

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

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

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

Divide both sides by $\sum_i \alpha_i$, so

- Set $\gamma_i = \frac{\alpha_i}{\sum_i^k \alpha_i}$.
- Thus, $\gamma_i \ge 0$ and $\sum_i \gamma_i = 1$

Convergence analysis

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

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

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

Divide both sides by $\sum_i \alpha_i$, so

- Set $\gamma_i = \frac{\alpha_i}{\sum_i^k \alpha_i}$.
- Thus, $\gamma_i \ge 0$ and $\sum_i \gamma_i = 1$

$$\mathbb{E}\left[\sum_i \gamma_i (F(x_i) - F(x^*))\right] \le \frac{r_1 + M^2 \sum_i \alpha_i^2}{2\sum_i \alpha_i}$$

Convergence analysis

- 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 $\gamma_i F(x_i)$

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**

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

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

$$\mathbb{E}\big[F(\bar{x}_k) - F(x^*)\big] \leq \frac{r_1 + M^2 \sum_i \alpha_i^2}{2\sum_i \alpha_i}.$$

SGD – finally

- Let $D_{\mathcal{X}} := \max_{x\in\mathcal{X}} \|x - x^*\|_2$ (act. only need $\|x_1 - x^*\| \le D_{\mathcal{X}}$)
- Assume $\alpha_i = \alpha$ is a constant. Observe that

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

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

Stochastic optimization – smooth

Theorem. Let $f(x,\xi)$ be C_L^1 convex. Let $e_k := \nabla F(x_k) - g_k$ satisfy $\mathbb{E}[e_k] = 0$. Let $\|x_i - x^*\| \le D$. Also, let $\alpha_i = 1/(L+\eta_i)$. Then,

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

Stochastic optimization – smooth

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 $\|x_i - x^*\| \le D$. Also, let $\alpha_i = 1/(L+\eta_i)$. Then,

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

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

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

Stochastic optimization – smooth

Theorem. Let $f(x,\xi)$ be C_L^1 convex. Let $e_k := \nabla F(x_k) - g_k$ satisfy $\mathbb{E}[e_k] = 0$. Let $\|x_i - x^*\| \le D$. Also, let $\alpha_i = 1/(L+\eta_i)$. Then,

$$\mathbb{E}\big[\sum\nolimits_{i=1}^k F(x_{i+1}) - F(x^*)\big] \le \frac{D^2}{2\alpha_k} + \sum\nolimits_{i=1}^k \frac{\mathbb{E}[\|e_i\|^2]}{2\eta_i}.$$

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

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

- Using $\alpha_i = L + \eta_i$ where $\eta_i \propto 1/\sqrt{i}$ we obtain
$$\mathbb{E}[F(\bar{x}_k) - F(x^*)] = O(\tfrac{LD^2}{k}) + O(\tfrac{\sigma D}{\sqrt{k}})$$

where σ bounds the variance $\mathbb{E}[\|e_i\|^2]$

Minimax optimal rate

Stochastic optimization – strongly convex

Theorem. Suppose $f(x,\xi)$ are convex and $F(x)$ is μ-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

Cvx Class	Rate	Iterate	Minimax
C_L^0	$1/\sqrt{k}$	$\bar{x}_k$	Yes
C_L^0	$\log k/\sqrt{k}$	x_k	No
C_L^1	$1/\sqrt{k}$	$\bar{x}_k$	Yes
S_L^0	$(\log k)/k$	$\bar{x}_k$, x_k	No
S_L^1	$1/k$	$\bar{x}_k$, x_k	Yes

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: $\widehat{F}(x) := \frac{1}{n}\sum_{i=1}^{n} f(x,\xi_i)$
- 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: $\widehat{F}(x) := \frac{1}{n}\sum_{i=1}^{n} f(x,\xi_i)$
- aka *Empirical Risk Minimization*
- **Note:** we often optimize $\widehat{F}$ using stochastic subgradient; but theoretical guarantees are then only on the *empirical* suboptimality $E[\widehat{F}(\bar{x}_k)] \leq \ldots$

SAA / Batch problem

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

Sample Average Approximation (SAA):

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

Finite-sum problems

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

Finite-sum problems

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

Gradient / subgradient methods

$$\begin{aligned}
x_{k+1} &= x_k - \alpha_k \nabla f(x_k) \\
x_{k+1} &= x_k - \alpha_k g(x_k), \quad g \in \partial f(x_k) \\
x_{k+1} &= \operatorname{prox}_{\alpha_k r}(x_k - \alpha_k \nabla f(x_k))
\end{aligned}$$

Stochastic gradient

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?

Stochastic gradient

Method	Assumptions	Full	Stochastic
Subgradient	convex	$O(1/\sqrt{k})$	$O(1/\sqrt{k})$
Subgradient	strongly cvx	$O(1/k)$	$O(1/k)$

So using stochastic subgradient, solve n times faster.

Stochastic gradient

Method	Assumptions	Full	Stochastic
Subgradient	convex	$O(1/\sqrt{k})$	$O(1/\sqrt{k})$
Subgradient	strongly cvx	$O(1/k)$	$O(1/k)$

So using stochastic subgradient, solve n times faster.

Method	Assumptions	Full	Stochastic
Gradient	convex	$O(1/k)$	$O(1/\sqrt{k})$
Gradient	strongly cvx	$O((1-\mu/L)^k)$	$O(1/k)$

– 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 ∇f_i for $i = 1, .., n$
- 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 ∇f_i for $i = 1, .., n$
- Select uniformly at random $i(k) \in \{1, \ldots, n\}$
- Perform the update

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

- Randomized / stochastic version of incremental gradient method of Blatt et al (2008)
- 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 $x^T a_i$)

SAG

Method	Assumptions	Rate
Gradient	convex	$O(1/k)$
Gradient	strongly cvx	$O((1-\mu/L)^k)$
Stochastic	strongly cvx	$O(1/k)$
SAG	strongly convex	$O((1-\min\{\frac{\mu}{n}, \frac{1}{8n}\})^k)$

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, . . .