DSC 291: Stochastic Optimization

What this course is about:

Most tasks in supervised learning can be modeled as

$$\text{(*)} \quad \min_{w} \; \mathop{\mathbb{E}}_{z \sim P} \ell(w, z) + \lambda \cdot r(w)$$

Here

- z is usually a feature/label pair $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}$
- $\ell(\cdot, z)$ is a random loss function

e.g.:
- $\ell(w, (x,y)) = (f_w(x) - y)^2$ MSE
- $\ell(w, (x,y)) = \log(1 + e^{-y \cdot w^T \phi(x)})$ Logistic

- $r: \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a structure inducing regularizer

e.g.:
- $r(w) = \|w\|_2^2$ Ridge
- $r(w) = \|w\|_1$ ℓ_1-penalty

Assumption: only access to P is through sampling $z \sim P$.

Two approaches:

Strategy 1: Emperical Risk Minimization

$$w_n = \arg\min_w \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i) + \lambda \cdot r(w)$$

where $z_1, \ldots, z_n \overset{iid}{\sim} P$

Strategy 2: Stochastic Approximation

Algorithm that produces a sequence of points $w_1, \ldots, w_t$, which in each iteration t draws a fresh batch

$$S_t = \{z_{t,1}, z_{t,2}, \ldots, z_{t,b}\} \overset{iid}{\sim} P$$

and uses $\{f(\cdot, z)\}_{z \in S}$ to update S.

Main Example: Stochastic Gradient Method

$$\left\{ \begin{array}{l} \text{Draw batch } S_t \\ w_{t+1} = w_t - \frac{\eta}{b} \sum_{z \in S} \nabla \ell(w_t, z) \end{array} \right\}$$

When $r = 0$, $\ell(\cdot, z)$ is smooth.

Goal: Solve using as few samples as possible.

Background:

(1.1) Inner products and linear maps.

- $\mathbb{E}$ is an Euclidean space: finite dimensional real vector space with an inner-product $\langle \cdot, \cdot \rangle : \mathbb{E} \times \mathbb{E} \to \mathbb{R}$ satisfying

(symmetry) $\langle x, y \rangle = \langle y, x \rangle$

(Bilinearity) $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$

$\forall x, y, z \in \mathbb{E}, \; a, b \in \mathbb{R}$

(Positive Semidefiniteness)

$\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$.

Ex: $\mathbb{E} = \mathbb{R}^d$ are arrays of d real values with

$$\langle x, y \rangle = x^T y = \sum_{i=1}^{d} x_i y_i$$

Ex: $\mathbb{E} = \mathbb{R}^d$ with an inner product

$$\langle x, y \rangle_A = \underset{1 \times d}{x^T} \underset{d \times d}{A} \underset{d \times 1}{y} = \sum_{i,j} A_{ij} x_i y_j,$$

where $A \in \mathbb{R}^{d \times d}$ is symmetric and positive definite (will come back to this in a second)

Ex: $\mathbb{E} = \mathbb{R}^{m \times n}$ with trace product

$$\langle X, Y \rangle := \langle \mathrm{vec}(X), \mathrm{vec}(Y) \rangle$$

$$= \sum_{i,j} X_{ij} Y_{ij}$$

$$= \mathrm{tr}(X^T Y)$$

check this

(1.2) Norms:

A norm on a vector space $\mathcal{V}$ is a function $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ satisfying

(absolute homogeneity) $\|ax\| = |a| \cdot \|x\|$

(triangle inequality) $\|x+y\| \leq \|x\| + \|y\|$

(Positivity) $\|x\| = 0 \iff x = 0$

$\forall x, y \in \mathcal{V}$ $a \in \mathbb{R}$.

Ex: If $\langle \cdot, \cdot \rangle$ is an inner product on V, then $\|x\| := \sqrt{\langle x, x \rangle}$ is a norm

e.g: $\mathbb{R}^n$ with dot-product $\rightarrow \|x\|_2 := \sqrt{\langle x, x \rangle} = \sqrt{x^T x} = \sqrt{\sum_{i=1}^n x_i^2}$

induced norm by inner-product.

Ex: $E = \mathbb{R}^{m \times n}$ with trace inner product.

Then $\|X\|_F := \sqrt{\langle X, X \rangle} = \sqrt{\mathrm{tr}(X^T X)}$

$= \|\mathrm{vec}(X)\|_2$

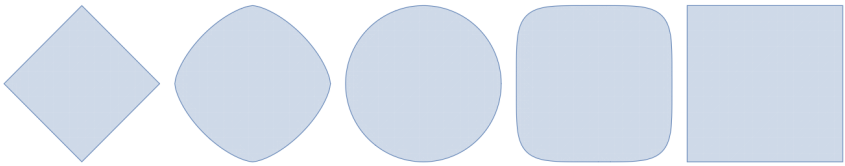
$= \sqrt{\sum_{i,j} X_{ij}^2}$

Frobenius norm

Lemma: (Cauchy-Schwartz)

$|\langle x, y \rangle| \leq \|x\| \cdot \|y\|$

induced norms



(a) $p = 1$ (b) $p = 1.5$ (c) $p = 2$ (d) $p = 5$ (e) $p = \infty$

Figure 1.1: Unit balls of ℓ_p-norms.

Ex: $\mathcal{V} = \mathbb{R}^d$. Define the ℓ_p-norm:

$$\|x\|_p := \begin{cases} \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}, & \text{if } 1 \le p < \infty \\ \max_{i=1,\dots,d} |x_i|, & \text{if } p = \infty \end{cases}$$

This is a norm. None of these are induced by any inner-product except $p=2$.

Thm: (von Neumann)
If $\varphi : \mathbb{R}^d \to \mathbb{R}$ is sign-permutation invariant norm on $\mathbb{R}^d$, then

$$\Phi(X) \triangleq \varphi(\sigma(X))$$

($\sigma(X)$: singular values; $X \in \mathbb{R}^{m \times n}$)

is a norm on $\mathbb{R}^{m \times n}$.

Eg:

$\|X\|_* = \|\sigma(X)\|_1$ Nuclear norm

$\|X\|_{op} = \|\sigma(X)\|_\infty$ Operator norm

$\|X\|_F = \sqrt{\mathrm{tr}(X^T X)} = \|\sigma(X)\|_2$

Fact: all norms on $\mathbb{E}$ are "equivalent"

For any two norms ρ_1 and ρ_2 on $\mathbb{E}$, there exists $\alpha, \beta > 0$ s.t.

$$\alpha \rho_1(x) \leq \rho_2(x) \leq \beta \rho_1(x) \quad \forall x \in \mathbb{E}$$

So ρ_1 and ρ_2 are within a constant multiplicative factor of each other.

Key caveat: α, β typically depend on the dimension $d = \dim(\mathbb{E})$

Ex: $\|x\|_2 \leq \|x\|_1 \leq \sqrt{d}\,\|x\|_2$

$\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d}\,\|x\|_\infty$

$\|x\|_\infty \leq \|x\|_1 \leq d\,\|x\|_\infty$.

We often think of d as being huge, so this is not super helpful.

(1.3) Differentiability

- E and Y are Euclidean spaces.
- $U \subseteq E$ be open set.
- A map $F: U \to Y$

is continuous at $x \in U$ if for any $x_i \xrightarrow[U]{} x$ we have $F(x_i) \to F(x)$

- F is L-Lipschitz if (Lipschitz constant)

$$\|F(y) - F(x)\| \leq L\|y - x\| \qquad \forall x, y \in Q.$$

If $L \in [0,1)$ we say that F is a contraction. If $L = 1$, we call F nonexpansive.

Defn: $f: E \to \mathbb{R}$ is differentiable at $x \in U$ if there exists $\nabla f(x) \in E$ s.t.

$$\lim_{h \to 0} \frac{f(x+h) - [f(x) + \langle \nabla f(x), h \rangle]}{\|h\|} = 0$$

We then call $\nabla f(x)$ the gradient.

Defn: $f: \mathbb{E} \to \mathbb{R}$ is twice differentiable at $x \in \mathbb{E}$ if there exists linear map $\nabla^2 f(x): \mathbb{E} \to \mathbb{E}$ s.t.

$$\lim_{h \to 0} \frac{\nabla f(x+h) - \nabla f(x) - \nabla^2 f(x) h}{\|h\|} = 0$$

We then call $\nabla^2 f(x)$ the Hessian.

Ex: $\mathbb{E} = \mathbb{R}^d$. Then

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_d} \end{bmatrix}$$

if f is differentiable

and $\nabla^2 f(x) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j} \right]_{i,j}$

if f is twice differentiable.

HW: compute $\nabla f(X)$, $\nabla^2 f(X)$ for

$$f(X) = \log \det(X).$$

Def: f is C^1-smooth if ∇f exists and is continuous.
f is C^2-smooth if $\nabla^2 f$ exists and is continuous.

Lemma (Accuracy in approx. motion)
If ∇f is L-Lipschitz, then

$$\left| f(y) - \underbrace{\left(f(x) + \langle \nabla f(x), y - x \rangle\right)}_{=: l_x(y)} \right| \leq \frac{L}{2} \|y - x\|^2$$

linearization.

pf: Homework.

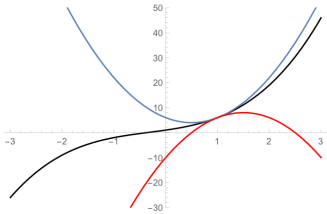


Figure 1.5: The black curve depicts the graph of a β-smooth function f; the blue and red curves depict graphs of the quadratics $l_x(\cdot) + \frac{\beta}{2}\|\cdot - x\|^2$ and $l_x(\cdot) - \frac{\beta}{2}\|\cdot - x\|^2$, respectively.

Thm: Consider C^2-smooth $f: \mathbb{R} \to \mathbb{R}$ and $x \in \mathbb{R}$. Then the following are true.

1) (Necessary) If $x \in \mathbb{R}$ is a local min. then $\nabla f(x) = 0$ and $\nabla^2 f(x) \geq 0$.

2) (Sufficient) If $\nabla f(x) = 0$ and $\nabla^2 f(x) > 0$ then x is a local minimizer of f.

1.4 Convexity

Defn: $Q \subseteq \mathbb{E}$ is convex if $\lambda x + (1-\lambda) y \in Q$ $\forall x, y \in \mathbb{E}$, $\lambda \in [0,1]$

Defn: $f: \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ is convex if

$$f(\lambda x + (1-\lambda) y) \leq \lambda \cdot f(x) + (1-\lambda) f(y) \quad \forall x, y \in \mathbb{E}, \ \lambda \in [0,1]$$

Thm: The following are equivalent for any C^1-smooth function $f: U \to \mathbb{R}$ defined on a convex open set $U \subseteq \mathbb{E}$.

(a) (convexity) f is convex:

(b) (gradient inequality)

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \forall x, y$$

If f is C^2-smooth, then also

(c) $\nabla^2 f(x) \succeq 0 \quad \forall x \in U$.

Note: For any C^1-smooth convex f, TFAE

- $\nabla f(x) = 0$
- x is a global minimizer

Chapter I: (Stochastic) Gradient Descent for linear least squares.

Problem:

(A) $$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{2n} \|Ax - b\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} (a_i^T x - b_i)^2$$

where $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, and a_i are the rows of A.

Ex: Learning from the diabetes data set. Goal: predict diabetes disease progression one year after baseline.

Data: $\{(x_i, b_i)\}_{i=1}^n$

- Number of samples: $n = 442$ patients
  - Training set (75%): 330
  - Validation set (25%): 110
- Number of features: $d = 10$

You will experiment with this dataset in your homework.

$f(x) = \frac{1}{2n}\|Ax-b\|_2^2$

Observe:

$\nabla f(x) = \frac{1}{n} A^T(Ax-b)$, $\nabla^2 f(x) = \frac{1}{n} A^T A$ (HW)

Notice: $\frac{1}{n} A^T A \succeq 0$ [Reason $v^T A^T A v = \|Av\|_2^2 \geq 0 \ \forall v$]

So f is convex.

So x is optimal for (★)

$\Rightarrow \frac{1}{n} A^T(Ax-b) = 0$

$\Leftrightarrow \boxed{A^T A x = A^T b}$

normal eqns.

Goal: Develop algorithms for solving normal eqns with cheap per iteration cost, and in particular that do not require forming $A^T A$.

$$\min_x \frac{1}{n}\|Ax-b\|_2^2 =: f(x)$$

Gradient Descent ($\eta > 0$)

$$x_{t+1} = x_t - \eta \nabla f(x_t) = x_t - \frac{\eta}{n}\left(A^T(Ax_t - b)\right)$$

Let $\bar{x}$ be a minimizer:

$$x_{t+1} - \bar{x} = x_t - \bar{x} - \frac{\eta}{n}\left(A^T(Ax_t - b)\right)$$

Recall $A^TA\bar{x} = A^Tb$

$$= x_t - \bar{x} - \frac{\eta}{n}\left(A^TAx_t - A^TA\bar{x}\right)$$

$$= \left(I - \frac{\eta}{n}A^TA\right)(x_t - \bar{x})$$

Summarizing

$$\boxed{x_{t+1} - \bar{x} = \left(I - \frac{\eta}{n}A^TA\right)(x_t - \bar{x})}$$

Define $H = \frac{1}{n} A^T A$ and $\Delta_t = x_t - \bar{x}$.

Then we have learned:

$$\Delta_t = (I - \eta H)\Delta_{t-1} = (I - \eta H)^t \Delta_0$$

Let $\alpha = \lambda_{min}(H)$, $\beta = \lambda_{max}(H)$, $\kappa = \frac{\beta}{\alpha}$.

Setting $\eta = \frac{1}{\beta}$ we see

$$\| I - \eta H \|_{op} = \max_{\lambda \in [\alpha, \beta]} |1 - \eta \lambda| = 1 - \kappa^{-1}$$

Therefore $\| \Delta_t \|_2 \leq (1 - \kappa^{-1})^t \| \Delta_0 \|_2$.

In particular using $1 - a \leq e^{-a}$ $\forall a \in [0,1]$, we may write

$$\| \Delta_t \|_2 \leq \exp\left(-\frac{t}{\kappa}\right) \| \Delta_0 \|_2 \leq \varepsilon$$

Solving for ε yields

$$\boxed{t \geq \kappa \log\left(\frac{\| \Delta_0 \|_2}{\varepsilon}\right)}$$

Let's now look instead at $f(x_t) - f(\bar{x})$ ($= f^*$)

Since f is a pure quadratic,

$$f(x) = f(\bar{x}) + \underbrace{\langle \nabla f(\bar{x}), x - \bar{x} \rangle}_{= 0} + \frac{1}{2} \underbrace{\langle \nabla^2 f(\bar{x})(x - \bar{x}), x - \bar{x} \rangle}_{= H}$$

Therefore

$$f(x_t) - f^* = \frac{1}{2} \Delta_0^T (I - \eta H)^{2t} H \Delta_0$$

A very similar argument as before shows

$$f(x_t) - f^* \leq (1 - \kappa^{-1})^{2t} (f(x_0) - f^*)$$

for $\eta = \frac{1}{\beta}$. In fact, we can obtain a rate that is insensitive to α:

$$f(x_t) - f^* \leq \frac{1}{2} \max_{\lambda \in [\alpha, \beta]} \lambda (1 - \tfrac{\lambda}{\beta})^{2t} \cdot \|\Delta_0\|_2^2$$

Observe

$$\lambda (1 - \tfrac{\lambda}{\beta})^{2t} \leq \lambda \exp(-\tfrac{2t\lambda}{\beta}) = \frac{\beta}{2t} \cdot \frac{2t\lambda}{\beta} \exp(-\tfrac{2t\lambda}{\beta})$$

$$\leq \frac{\beta}{2t} \cdot \underbrace{\max_{s > 0} s e^{-s}}_{= e^{-1}}$$

Thus $f(x_t) - f^* \leq \frac{\beta \|\Delta_0\|^2}{8t} \leq \varepsilon$

Solving for $\varepsilon > 0$ gives

$$t \geq \frac{\beta \|\Delta_0\|^2}{8\varepsilon}$$

Summary: Gradient Descent with $\eta = \frac{1}{\beta}$ enjoys the guarantee:

$$f(x_t) - f^* \leq \min\left\{ (1 - \kappa^{-1})^{2t} (f(x_0) - f^*), \frac{\beta \|x_0 - \bar{x}\|_2^2}{t} \right\}$$

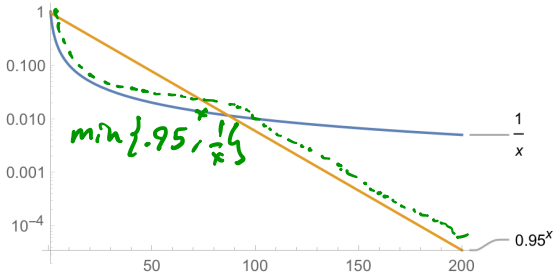


Figure 5.2: Sublinear vs. linear rates

Chapter 2:

Stochastic gradient descent:

Each iteration of GD requires $n \cdot d$ arithmetic operations. SGD has cost $O(d)$ per iteration instead.

$$\min_x f(x) := \mathbb{E}_{(a,b)\sim p} \frac{1}{2}(\langle a, x\rangle - b)^2$$

SGD: step $t = 1, \dots, T$:

- Draw $(a_t, b_t) \sim p$
- $x_{t+1} = x_t - \eta_t (\langle a_t, x_t\rangle - b_t) \cdot a_t$

Define

$$\Sigma := \mathbb{E}[aa^T], \quad v = \mathbb{E}[ba], \quad \bar{x} \in \operatorname{argmin} f.$$

Note

$$0 = \nabla f(\bar{x}) = \mathbb{E}(\langle a, \bar{x}\rangle - b)a = \mathbb{E}[aa^T\bar{x} - ba]$$

$$\Rightarrow \boxed{\Sigma \bar{x} = v}$$

Set

$$\Delta_t = x_t - \bar{x}, \qquad \varepsilon_t = b_t - \langle a_t, \bar{x} \rangle$$

So $\langle a_t, x_t \rangle - b_t = \langle a_t, \Delta_t \rangle - \varepsilon_t$

Therefore

$$\Delta_{t+1} = \Delta_t - \eta_t (\langle a_t, \Delta_t \rangle - \varepsilon_t) a_t$$

$$= \left(I - \eta_t a_t a_t^T\right) \Delta_t + \eta_t \varepsilon_t a_t$$

and as in the deterministic case

$$f(x) - f^* = \frac{1}{2} (x - \bar{x})^T \Sigma (x - \bar{x})$$

Let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid x_t]$

Assumptions: ← well-specified

- $\mathbb{E}[\varepsilon | a] = 0$, $\mathbb{E}[\varepsilon^2 | a] \le \sigma^2$,

  $\alpha I \le \Sigma$, $\mathbb{E}\left[\|a\|^2 a a^T\right] \le R^2 \Sigma$

  e.g. true for $a \sim N(0, \Sigma)$

  e.g. true for $\|a\| \le R$

Write

$$\|\Delta_{t+1}\|_2^2 = \underbrace{\|(I-\eta_t a_t a_t^T)\Delta_t\|_2^2}_{A} \underbrace{- 2\eta_t \langle \xi_t a_t, (I-\eta_t a_t a_t^T)\Delta_t \rangle}_{B} + \underbrace{\eta_t^2 \|\xi_t a_t\|_2^2}_{C}$$

$$\mathbb{E}_t A = \Delta_t^T \mathbb{E}(I-\eta_t a_t a_t^T)^2 \Delta_t$$

$$= \Delta_t^T \left(I - 2\eta_t \Sigma + \eta_t^2 \mathbb{E}[\|a\|_2^2 \cdot a a^T] \right) \Delta_t$$

$$\leq \Delta_t^T \left(I - \eta_t (2 - \eta_t R^2)\Sigma \right) \Delta_t$$

$$\mathbb{E}_t B = 2\eta_t \langle \underbrace{\mathbb{E}[\xi_t | a_t]}_{=0} \cdot a_t, \dots \rangle = 0$$

$$\mathbb{E}_t C = \eta_t^2 \mathbb{E}_{a_t}\left[\mathbb{E}[\xi_t^2 | a_t] \cdot \|a_t\|^2\right] \leq \eta_t^2 \sigma^2 \underbrace{\mathbb{E}\|a_t\|^2}_{\text{tr}(\Sigma)}$$

So

$$\mathbb{E}_t \|\Delta_{t+1}\|_2^2 \leq \left(1 - \alpha\eta_t(2-\eta_t R^2)\right)\|\Delta_t\|_2^2 + \eta_t^2 \sigma^2 \text{tr}(\Sigma)$$

Take expectation over $x_1, \dots, x_t$ and use tower rule to get:

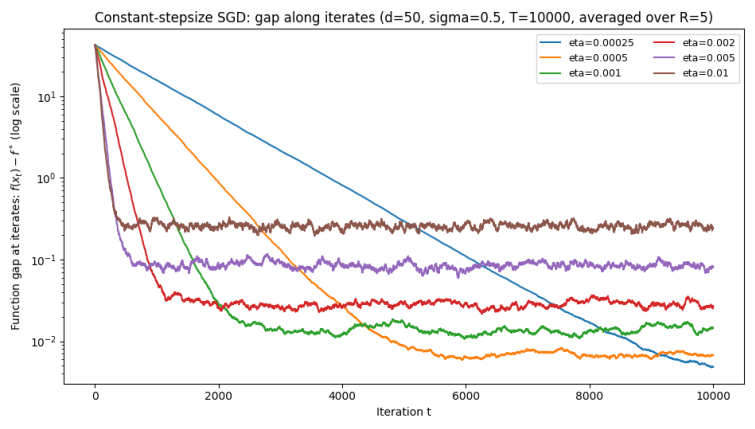
$:= d_t$

$$\textcircled{\star}\quad \underbrace{\mathbb{E}\|\Delta_{t+1}\|_2^2}_{=:d_{t+1}} \le (1-\alpha\eta_t(2-\eta_t R^2))\,\underbrace{\mathbb{E}\|\Delta_t\|^2}_{:=d_t} + \eta_t^2\sigma^2\cdot tr(\Sigma)$$

Conclusion:

With constant stepsize $\eta \le \frac{1}{R^2}$, get

$$\begin{aligned} d_{t+1} &\le (1-\alpha\eta)\,d_t + \eta^2\sigma^2 tr(\Sigma) \\ &\vdots \\ &\le (1-\alpha\eta)^t d_0 + \eta^2\sigma^2 tr(\Sigma)\cdot\sum_{i=0}^{t}(1-\alpha\eta)^i \\ &\le \underbrace{(1-\alpha\eta)^t d_0}_{\text{initial conditions decay exponentially fast}} + \underbrace{\frac{\eta\sigma^2 tr(\Sigma)}{\alpha}}_{\text{noise floor}} \end{aligned}$$



To get to zero error. there are two strategies

Thm: Set $\eta_t = \frac{2}{\alpha t + 2R^2}$. Then x_t satisfy

$$d_t \le \frac{\max\left\{\alpha^2\left(1+\frac{2R^2}{\alpha}\right)d_0,\ 4\sigma^2 \cdot tr(\Sigma)\right\}}{\alpha^2\left(t+\frac{2R^2}{\alpha}\right)}$$

[So roughly $d_t = O\left(\frac{\sigma^2 \cdot tr(\Sigma)}{\alpha^2 t}\right)$]

pf: From ⊛:

$$d_{t+1} \le \left(1 - \frac{2}{t + 2R^2/\alpha}\right) d_t + \frac{4\sigma^2 \cdot tr(\Sigma)/\alpha^2}{(t + 2R^2/\alpha)^2}.$$

The rest follows from the following lemma, which you will prove for HW.

Lemma: Consider sequence $D_t > 0$ and constants $t_0 \ge 0$, $a > 0$ satisfying

$$D_{t+1} \le \left(1 - \frac{2}{t+t_0}\right) D_t + \frac{a}{(t+t_0)^2}$$

Then $D_t \le \frac{\max\{(1+t_0)D_1,\ a\}}{t+t_0}$ $\forall t$.

□

Second strategy is to run SGD in epochs

Recall for fixed t and $\eta \equiv \eta_0 = \frac{1}{L^2}$:

$$d_{t+1} \le (1-\alpha\eta_0)^t d_0 + \eta_0 \frac{\sigma^2 \mathrm{tr}(\Sigma)}{\alpha}$$

So after $t \ge \frac{1}{\alpha\eta_0} \log\left(\frac{d_0}{2\eta_0\sigma^2\mathrm{tr}(\Sigma)/\alpha}\right)$

can be sure $d_{t+1} \le \frac{2\eta_0\sigma^2\mathrm{tr}(\Sigma)}{\alpha}$

Now replace η_0 by $\eta_1 = \eta_0/2$ and repeat: In next epoch, can be sure that after

$$t \ge \frac{1}{\alpha\eta_1} \log\left(\frac{2\eta_0\sigma^2\mathrm{tr}(\Sigma)/\alpha}{2\eta_1\sigma^2\mathrm{tr}(\Sigma)/\alpha}\right)$$

$$= \frac{1}{\alpha\eta_1}\log(2)$$

can be sure $d_{t+1} \le \frac{2\eta_1\sigma^2\mathrm{tr}(\Sigma)}{\alpha}$

You only need repeat this until

$$2^{-s}\eta_0 = \eta_s \le \frac{\varepsilon\alpha}{2\sigma^2\mathrm{tr}(\Sigma)}$$ to get $d \le \varepsilon$.

$$\Rightarrow s \le \log\left(\frac{1}{\varepsilon}\right)$$

So total # iterations after careful accounting:

$$\frac{R^2}{\alpha}\log\left(\frac{d_0}{\epsilon}\right) + \frac{2\sigma^2 \operatorname{tr}(\Sigma)}{\epsilon\alpha^2}$$

+ obtain a point that is ϵ-close to $\bar{x}$.

Remarkable Fact: the dependence on α can be removed!!! when tracking the function values $f(\hat{x}_t) - f^*$ along the average iterate:

$$\hat{x}_t := \frac{1}{t+1}\sum_{i=0}^{t} x_i$$

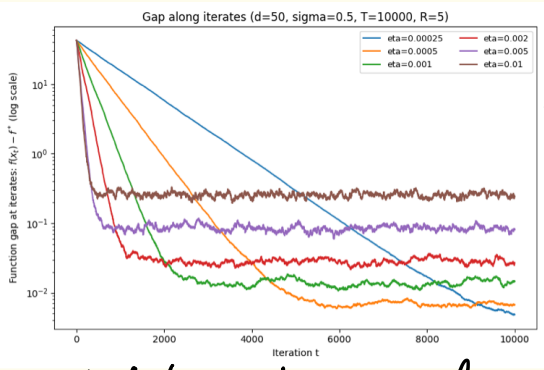
Thm: Set $\alpha = \lambda_{\min}(\Sigma)$ $\beta = \lambda_{\max}(\Sigma)$ and $\eta = \frac{1}{\beta}$ (constant!).

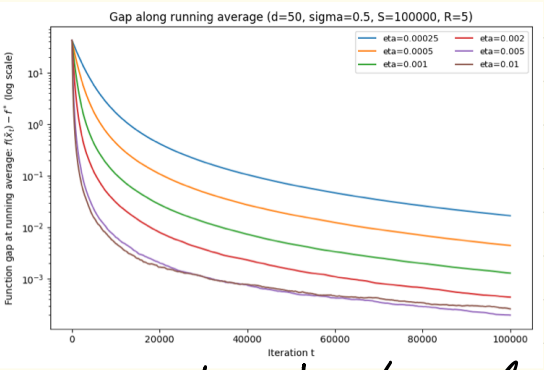
Suppose $\mathbb{E}[\varepsilon | a] = 0$. Then it holds:

$$\mathbb{E}\left(f(\hat{x}_T) - f^*\right] \le \frac{\beta^2 \cdot \Delta_0^T \Sigma^{-1} \Delta_0}{2T^2} + \frac{\sigma^2 d}{2T} + O\left(\frac{\beta d_0}{T^2} + \frac{1}{\beta T}\right)$$

low-order

↳ See Bach-Moulines 2013





Pf sketch: The proof is quite technical and we will only sketch it. We have seen that Δ_t evolve according to

$$\Delta_{t+1} = (I - \eta a_t a_t^T)\Delta_t + \eta \varepsilon_t a_t$$

$$= (I - \eta \Sigma)\Delta_t + \eta \underbrace{(\varepsilon_t a_t - (a_t a_t^T - \Sigma)\Delta_t)}_{Q}$$

Observe $\mathbb{E}_t Q^2 = O(\|\Delta_t\|_2^2) = O(\eta)$

so Q^2 should be dominated by $\varepsilon_t a_t$. Making this precize requires work. But taking it for granted let's analyze the evolution

$$\Delta_{t+1} = (I - \eta\Sigma)\Delta_t + \eta q_t \quad \text{with } q_t = \varepsilon_t a_t$$

Let's expand the recursion:

$$\Delta_t = (I-\eta\Sigma)^t \Delta_0 + \eta \sum_{i=0}^{t} (I-\eta\Sigma)^i q_{t-i}$$

$$\mathbb{E}\|\cdot\|_2^2 \approx \eta\sigma^2 \mathrm{tr}(\Sigma)$$

(same computation as before)

Now let's look at averaging:

$$\hat{x}_t - \bar{x} = \frac{1}{t+1}\sum_{i=0}^{t}(x_i - \bar{x}) = \frac{1}{t+1}\sum_{i=0}^{t}\Delta_i$$

Therefore

$$\hat{x}_{t-1} - \bar{x} = \frac{1}{t}\sum_{i=0}^{t-1}\Delta_i$$

$$= \underbrace{\frac{1}{t}\sum_{j=0}^{t-1}(I-\eta\Sigma)^j \Delta_0}_{B} + \underbrace{\frac{\eta}{t}\sum_{j=0}^{t-1}\sum_{i=0}^{j}(I-\eta\Sigma)^i q_{j-i}}_{N}$$

B: bias term

N: noise term

Observe

$$\sum_{j=0}^{t-1}(I-\eta\Sigma)^j = (\eta\Sigma)^{-1}\cdot\left(I-(I-\eta\Sigma)^t\right)$$

↑ geometric series.

So $B = \frac{1}{t}(\eta\Sigma)^{-1}\cdot\left(I-(I-\eta\Sigma)^t\right)\Delta_0$.

For the noise term, swap sums:

$$N = \frac{\eta}{t}\sum_{j=0}^{t-1}\sum_{i=0}^{j}(I-\eta\Sigma)^i q_{j-i}$$

$$= \frac{\eta}{t}\sum_{k=0}^{t-1}\underbrace{\left(\sum_{s=0}^{t-1-k}(I-\eta\Sigma)^s\right)}_{\|}q_k$$

$$(\eta\Sigma)^{-1}\left(I-(I-\eta\Sigma)^{t-k}\right)$$

Now observe the algebraic identity

$$f(x)-f^* = \mathbb{E}(\langle a,x\rangle - b)^2 - f^*$$

$$= \mathbb{E}(\langle a,x\rangle - \langle a,\bar{x}\rangle - \varepsilon)^2 - \mathbb{E}\varepsilon^2$$

$$= \mathbb{E}\langle a, x-\bar{x}\rangle^2$$

$$= \mathbb{E}\,\mathrm{tr}\left(aa^T,(x-\bar{x})(x-\bar{x})^T\right)$$

$$= \underbrace{(x-\bar{x})^T\Sigma(x-\bar{x})}_{\|}$$

$$\|x-\bar{x}\|_\Sigma^2$$

So

$$f(\hat{x}_t) - f^* = \|\hat{x}_t - \bar{x}\|_\Sigma^2 = \|B+N\|_\Sigma^2$$

$$\leq 2\|B\|_\Sigma^2 + 2\|N\|_\Sigma^2$$

Now

$$\|B\|_\Sigma^2 = \frac{1}{\eta^2 T^2} \Delta_0^T \underbrace{\left(I-(I-\eta\Sigma)^t\right)\Sigma^{-1}\left(I-(I-\eta\Sigma)^t\right)}_{\lesssim \Sigma^{-1}} \Delta_0$$

$$\leq \frac{\Delta_0^T \Sigma^{-1} \Delta_0}{\eta^2 T^2}$$

Next define $A_{t,k} = I - (I-\eta\Sigma)^{t-k}$

Then

$$\mathbb{E}\|N\|_\Sigma^2 = \mathbb{E}\, N^T \Sigma N$$

$$= \frac{\eta^2}{t^2} \sum_{k=0}^{T-1} \mathbb{E}\Big[q_k^T \underbrace{A_{t,k}\Sigma^{-1}A_{t,k}}_{\lesssim \Sigma^{-1}} q_k\Big]$$

$$= \frac{\eta^2}{t^2} \sum_{k=0}^{T-1} \mathrm{tr}\left(\Sigma^{-1} \underbrace{\mathbb{E} q_k q_k^T}_{\sigma^2\Sigma}\right) \leq \frac{\eta^2\sigma^2 d}{t}$$

The role of mini-batches. In practice, you run mini-batch SGD:

$$\begin{cases} \text{Sample } S_t = \{(x_1, b_1), (x_2, b_2), \ldots, (x_s, b_s)\} \overset{iid}{\sim} p \\ x_{t+1} = x_t - \dfrac{\eta_t}{s} \underbrace{\sum_{(x,y) \in S_t} (a^T x - b) a}_{= g_t} \end{cases}$$

Advantage:

- For reasonably sized s, the gradients for each data point in the batch can be computed in parallel.
- Does it help? Yes! Let's look at the last iterate convergence

$$\Delta_{t+1} = \Delta_t - \eta_t g_t = \Delta_t - \eta_t \frac{1}{s} \sum_{(x,y)} (a^T x - b) a$$

$$= \left(I - \eta_t \underbrace{\frac{1}{s} \sum_{S} a a^T}_{\mathbb{E} = \Sigma \text{ same as here}}\right) \Delta_t + \eta_t \frac{1}{s} \sum_{S} \varepsilon a$$

What is

$$\mathbb{E}\left\| \frac{1}{S}\sum_{i=1}^{S} \epsilon_i a_i \right\|^2 = \frac{1}{S^2}\sum_{i=1}^{S} \mathbb{E}\left[\epsilon_i^2 \|a_i\|_2^2\right]$$

↑ analogue of $\sigma^2 \mathrm{tr}(\Sigma)$

$$= \frac{1}{S^2}\sum_{i=1}^{S} \sigma^2 \mathrm{tr}(\Sigma_i)$$

$$= \frac{1}{S}\cdot\sigma^2\cdot\mathrm{tr}(\Sigma_i)$$

All the results we have seen extend verbatim with σ^2 is replaced by $\frac{\sigma^2}{S}$.

For example, for the last iterate with $\eta_t \equiv \eta$, we get

$$\|\Delta_t\|_2^2 \le (1-\eta\alpha)^t \|\Delta_0\|_2 + \frac{\eta\sigma^2}{\alpha S}\mathrm{tr}(\Sigma)$$

In particular, if you do step-decay, you need

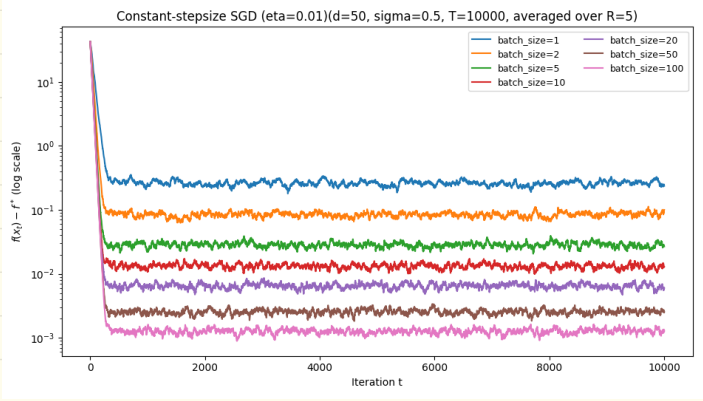
$$t \ge \frac{R^2}{\alpha}\log\left(\frac{\Delta_0}{\varepsilon}\right) + \frac{\sigma^2 \mathrm{tr}(\Sigma_i)}{S\alpha^2\varepsilon}$$

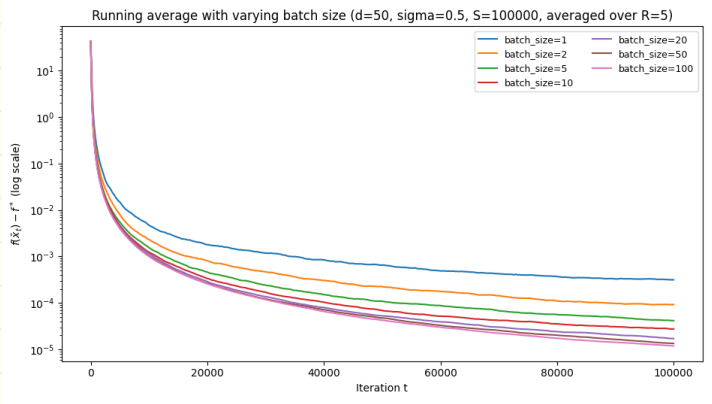
(bias / noise)

steps to ensure $\mathbb{E}\|x_t - \bar{x}\|^2 \le \varepsilon$.

Batch saturation:

For sufficiently large s and finite time horizon T, the noise term is smaller than the bias so no improvement is seen with larger batches.



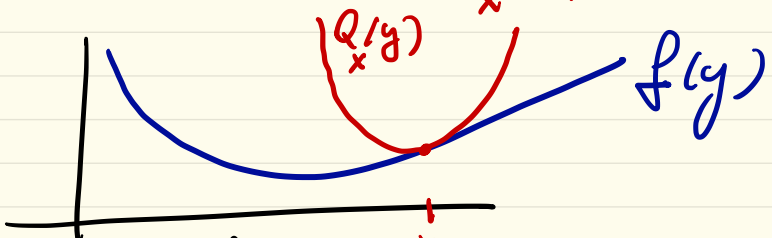


Chapter 2: General Convex Optimization

$$\min_x f(x)$$

where $f: \mathbb{E} \to \mathbb{R}$ is β-smooth:

$$f(y) \le \underbrace{f(x) + \langle \nabla f(x), y-x \rangle + \frac{\beta}{2}\|y-x\|^2}_{Q_x(y)} \quad \forall x,y \qquad (\star)$$



Remark: Follows if $\|\nabla f(x) - \nabla f(y)\| \le \beta \cdot \|x-y\|$

Gradient Descent:

$$x_{t+1} = x_t - \eta \nabla f(x_t) = \arg\min_x f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta}\|x - x_t\|_2^2$$

Lemma (Descent)

$$f(x_t) - f(x_{t+1}) \ge \eta\left(1 - \frac{\eta\beta}{2}\right)\|\nabla f(x_t)\|^2$$

pf: Plug in x_{t+1} into $(\star)$.

Since $\eta \mapsto \eta\left(1-\frac{\eta\beta}{2}\right)$ is maximized at $\eta=\frac{1}{\beta}$ we deduce that $x^+ = x - \frac{1}{\beta}\nabla f(x)$ satisfies

$$\boxed{f(x) - f(x^+) \geq \frac{1}{2\beta}\|\nabla f(x)\|^2}$$

Thm:

$$\min_{i=1,\dots,t} \|\nabla f(x_i)\|^2 \leq \frac{1}{t}\sum_{i=1}^{t}\|\nabla f(x_i)\|^2 \leq \frac{2\beta(f(x_1)-f^*)}{t}$$

Let's make stronger assumptions.

Thm: Let $f:\mathbb{E}\to\mathbb{R}$ be convex and β-smooth. Then with $\eta=\frac{1}{\beta}$, GD iterates satisfy:

$$\boxed{f(x_t) - f^* \leq \frac{\beta\|x_0-\bar{x}\|^2}{2t}}$$

Pf: Letting $Q_x(y) = f(x) + \langle \nabla f(x), y-x\rangle + \frac{\beta}{2}\|y-x\|^2$ get

$$\begin{aligned} f(x_{t+1}) &\leq Q_{x_t}(x_{t+1}) \\ &= f(x_t) + \langle \nabla f(x_t), x_{t+1}-x_t\rangle + \frac{\beta}{2}\|x_{t+1}-x_t\|^2 \\ &= f(x_t) + \langle \nabla f(x_t), \bar{x}-x_t\rangle \\ &\quad - \langle \nabla f(x_t), \bar{x}-x_{t+1}\rangle + \frac{\beta}{2}\|x_{t+1}-x_t\|^2 \\ &\leq f^* + \frac{\beta}{2}\left(\|x_t-\bar{x}\|^2 - \|x_{t+1}-\bar{x}\|^2\right) \end{aligned}$$

(cvx: $f(x_t) \leq f(\bar{x}) = f^*$)

Sum up these inequalities for $i = 1, \dots, t$.
Right-side telescopes:

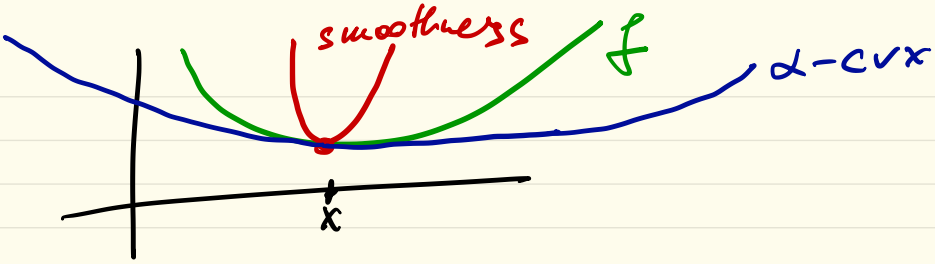
$$\underbrace{\frac{1}{t}\sum_{i=0}^{t-1}\left(f(x_i) - f^*\right)}_{\geq \min_{i=0,\dots,t-1} f(x_i) - f^*} \leq \frac{\beta \|x_0 - \bar{x}\|^2}{t-1}$$

$$\geq f(x_{t-1}) - f^*$$

Defn: f is called α-strongly cvx if $x \mapsto f(x) - \frac{\alpha}{2}\|x\|^2$ is still convex.

Lemma: Suppose f is C^1-smooth. Then f is α-strongly cvx iff

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|^2 \quad \forall x, y$$



Lemma: Suppose f is C^2-smooth. Then f is α-cvx iff $\nabla^2 f(x) \succeq \alpha \cdot I \quad \forall x$

For a β-smooth and α-cvx f define

condition # to be $\kappa = \frac{\beta}{\alpha}$

Lemma: (Polyak-Łojasiewicz) not needed with a diff proof

Let f be C^1-smooth and α-cvx.

Then

$$\frac{1}{2\alpha} \|\nabla f(x)\|^2 \geq f(x) - f^* \quad \forall x$$

pf: Define $Q_x(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y-x\|^2$

Then $f(\bar{x}) \geq Q_x(\bar{x}) \geq \min_y Q_x(y) = f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|^2$

□

Thm: Let f be β-smooth and α-cvx. Then GD iterates with $\eta = \frac{1}{\beta}$ satisfy.

① $f(x_{t+1}) - f^* \leq \left(1 - \frac{1}{2\chi}\right)(f(x_t) - f^*)$

② $\|x_{t+1} - \bar{x}\|^2 \leq \left(\frac{\chi - 1}{\chi + 1}\right)\|x_t - \bar{x}\|^2$

pf: We'll prove ① and leave ② for you to look at.

$$(f(x_t) - f^*) - (f(x_{t+1}) - f^*) \geq \frac{1}{2\beta}\|\nabla f(x_t)\|^2 \quad \text{(Descent lemma)}$$

$$\overset{PL}{\geq} \frac{\alpha}{2\beta}(f(x_t) - f^*)$$

$$\Rightarrow f(x_{t+1}) - f^* \leq \left(1 - \frac{1}{2\chi}\right)(f(x_t) - f^*)$$

Conclusion:

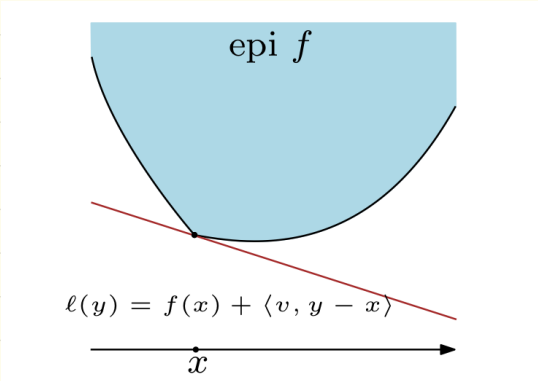
$$f(x_t) - f^* \leq \min\left\{\frac{\beta \cdot \|x_0 - \bar{x}\|^2}{2t}, \left(1 - \frac{1}{2\chi}\right)^t (f(x_0) - f^*)\right\}$$

Stochastic algorithms (basic ones) are almost unaffected by smoothness or lack thereof of the convex f. We will now focus on subgradients of nonsmooth CVX functions.

Defn: Let $f: \mathbb{E} \to \mathbb{R} \cup \{+\infty\}$ be CVX and let $x \in \mathbb{E}$ be s.t. $f(x) < \infty$. Then $v \in \mathbb{E}$ is a subgradient of f at x if

$$f(y) \geq f(x) + \langle v, y - x \rangle \quad \forall x, y$$

Then set of all such v is denoted by $\partial f(x)$ and is called the subdifferential



(a) Subgradient of a convex function

Properties: Let f, g be cvx, $A \in \mathbb{R}^{m \times n}$

- If f is C^1 at x, then
$$\partial f(x) = \{\nabla f(x)\}$$
- If f is finite near x, then

(Sum) $$\partial(f+g)(x) = \partial f(x) + \partial g(x)$$

where
$$A + B = \{a+b : a \in A,\ b \in B\}$$

- For $\varphi(x) = f(Ax-b)$, we have
$$\partial \varphi(x) = A^T \partial f(Ax-b)$$
provided f is finite near $Ax-b$.

Ex: $\varphi(x) = \|Ax-b\|_1 + \|x\|_\infty$

Then
$$\begin{aligned}\partial \varphi(x) &= \partial(x \mapsto \|Ax-b\|_1) + \partial \|\cdot\|_\infty(x) \\ &= A^T \partial \|\cdot\|_1 (Ax-b) + \partial \|\cdot\|_\infty(x)\end{aligned}$$

More rules:

- If $\varphi(x) = \sum_{i=1}^{d} \varphi_i(x_i)$, then

$$\partial \varphi(x) = \partial \varphi_1(x_1) \times \cdots \times \partial \varphi_d(x_d)$$

- If $f_1, \ldots, f_d$ are finite CVX and $\varphi(x) = \max_{i=1,\ldots,d} f_i(x)$. Then

$$\partial \varphi(x) = \mathrm{conv}\left\{ \bigcup_{i=I(x)} \partial f_i(x) \right\}$$

where $I(x) := \{i : f_i(x) = \varphi(x)\}$

So

$$\partial \|\cdot\|_1(x) = \partial|\cdot|(x_1) \times \cdots \times \partial|\cdot|(x_d)$$

$$\partial|\cdot|(s) = \begin{cases} \operatorname{sign}(s), & s \neq 0 \\ [-1,1], & s = 0 \end{cases}$$

And

$$\partial \|\cdot\|_\infty(x) = \operatorname{conv}\{A_i : i \in I(x)\}$$

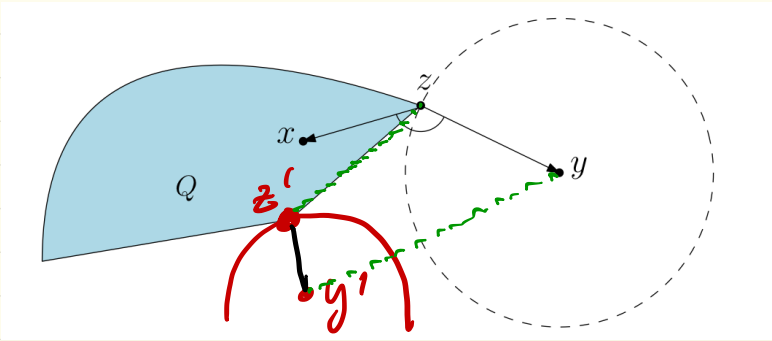
where $I(x) = \{i : |x_i| = \|x\|_\infty\}$

and $A_i = \{\mathbb{R}^q \times \mathbb{R}^d \times \cdots \times \underset{i\text{th entry}}{\partial|\cdot|(x_i)} \times \cdots \mathbb{R}^q\}$

Last theory bit:

For a set Q, we define

$$\text{proj}_Q(y) = \underset{x \in Q}{\text{argmin}} \; \|x - y\|_2^2$$



Thm: Suppose Q is closed cvx. Then $\text{proj}_Q(y)$ is a singleton for any y and satisfies

$$\|\text{proj}_Q(y') - \text{proj}_Q(y)\| \leq \|y' - y\| \quad \forall y, y'.$$

Problem.

$$\min_{x \in Q} f(x)$$

where Q is closed cvx and f is L-Lipschitz on a neighborhood of Q.

Ex: $\min_z \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i \langle \varphi(x_i), z \rangle\}$

Kernel Support Vector Machines

Ex: $\min \|Ax - b\|_1$
s.t. $\|x\|_p \leq \lambda$

Ex: $\min \|Ax - b\|_1$
s.t. $\sum_{i=1}^{d} x_i = 1, \quad x \geq 0$

Stochastic oracle: We assume existence of a map $(x,z) \mapsto g(x,z)$ with $z \sim P$ random s.t.

$$\mathbb{E}_z g(x,z) \in \partial f(x)$$

$$\mathbb{E}_z \|g(x,z)\|^2 \leq L^2 \quad \forall x \in Q$$

Ex: $g(x,z) = \nabla \ell(x,z)$, if $f(x) = \mathbb{E}\,\ell(x,z)$ if ℓ is smooth.

Ex: $g(x,z) \in \partial \ell(x,z)$, if $f(x) = \mathbb{E}\,\ell(x,z)$ and ℓ may be nonsmooth.

Projected stochastic gradient method:

$$\left\{ \begin{array}{l} \text{Draw } z_t \sim P \\ \text{Set } x_{t+1} = \text{proj}_Q\left(x_t - \eta_t g(x_t, z_t)\right) \end{array} \right\}$$

Thm: Suppose f is cvx and L-Lip on a neighborhood of Q. Then

$$\mathbb{E} f\left(\frac{1}{\sum_{i=0}^{t}\eta_i}\sum_{i=0}^{t}\eta_i x_i\right) - f^* \le \frac{\|x_0-\bar{x}\|^2 + L^2\sum_{i=0}^{t}\eta_i^2}{2\sum_{i=0}^{t}\eta_i}$$

In particular if $\eta_t = \frac{R}{L\sqrt{T+1}}$ where $R \ge \|x_0-\bar{x}\|$, we get

$$\mathbb{E} f\left(\frac{1}{T+1}\sum_{t=0}^{T} x_t\right) - f^* \le \frac{RL}{\sqrt{T+1}}$$

pf: Set $v_t = G(x_t, z_t)$. Then

$$\|x_{t+1}-\bar{x}\|^2 = \|\text{proj}_Q(x_t - \eta_t v_t) - \bar{x}\|^2$$
$$\le \|x_t - \eta_t v_t - \bar{x}\|^2$$
$$= \|x_t-\bar{x}\|^2 - 2\eta_t\langle v_t, x_t-\bar{x}\rangle + \eta_t^2\|v_t\|^2$$

We take $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid x_t]$ to get $v_t \in \partial f(x_t)$

$$\mathbb{E}_t\|x_{t+1}-\bar{x}\|^2 \le \|x_t-\bar{x}\|^2 - 2\eta_t\langle \mathbb{E}v_t, x_t-\bar{x}\rangle + \eta_t^2 L^2$$
$$\le \|x_t-\bar{x}\|^2 - 2\eta_t(f(x_t)-f^*) + \eta_t^2 L^2$$

Now take $\mathbb{E}$ w.r.t x_t to get

$$\mathbb{E}\|x_{t+1}-\bar{x}\|^2 \le \mathbb{E}\|x_t-\bar{x}\|^2 - 2\eta_t\left(\mathbb{E}f(x_t)-f^*\right)+\eta_t^2L^2$$

$$\vdots$$

$$\le \|x_0-\bar{x}\|^2 - 2\sum_{t=0}^{T}\eta_t\left(\mathbb{E}f(x_t)-f^*\right)+L^2\sum_{t=0}^{T}\eta_t^2$$

$$\Rightarrow \frac{\sum_{t=0}^{T}\eta_t\left(\mathbb{E}f(x_t)-f^*\right)}{\sum\eta_t} \le \frac{\|x_0-\bar{x}\|^2+L^2\sum_{t=0}^{T}\eta_t^2}{2\sum_{t=0}^{T}\eta_t}$$

$$\Downarrow$$

$$f(\text{average})-f^* \qquad \blacksquare$$

So to ensure $\mathbb{E}f(\hat{x})-f^*\le\varepsilon$ suffices

$$T \ge \frac{R^2L^2}{\varepsilon^2}$$

Note this is a factor ε worse than GD for smooth deterministic problems!

Thm: Suppose f is α-cvx and L-Lip on a neighborhood of closed cvx Q. Then with $\eta_t = \frac{2}{\alpha(t+1)}$ get

$$\mathbb{E} f\left(\frac{2}{t(t+1)} \sum_{i=1}^{t} i x_i\right) - f^* \le \frac{2L^2}{\alpha(t+1)}$$

pf: Same as before:

$$\mathbb{E}_t \|x_{t+1} - \bar{x}\|^2 \le \|x_t - \bar{x}\|^2 + 2\eta_t (f^* - f(x_t)) - \frac{\alpha}{2}\|x_t - \bar{x}\|^2) + \eta_t^2 L^2$$

Rearrange, tak $\mathbb{E}$ w.r.t. x_t, divide by η_t:

$$\mathbb{E} f(x_t) - f^* \le \left(\frac{1 - \alpha\eta_t}{2\eta_t}\right) \mathbb{E}\|x_t - \bar{x}\|^2 - \frac{1}{2\eta_t} \mathbb{E}\|x_{t+1} - \bar{x}\|^2 + \frac{\eta_t}{2} L^2$$

Algebraic trick: $\eta_t = \frac{2}{\alpha(t+1)}$ and multiply by t:

$$t\,\mathbb{E} f(x_t) - f^* \le \frac{\alpha t(t-1)}{4} \mathbb{E}\|x_t - \bar{x}\|^2 - \frac{\alpha t(t+1)}{4} \mathbb{E}\|x_{t+1} - \bar{x}\|^2 + \frac{t}{\alpha(t+1)} L^2$$

Sum up and telescope:

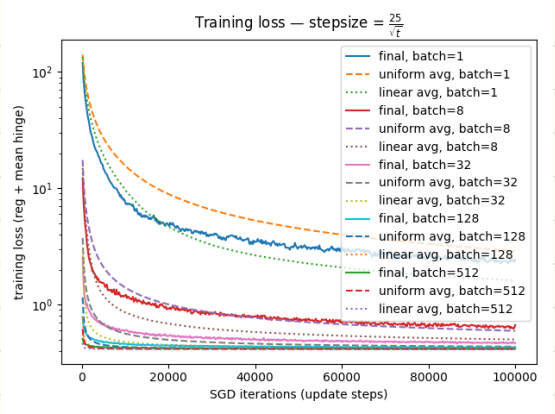
$$\sum_{i=1}^{t} i\,(\mathbb{E} f(x_i) - f^*) \le \sum_{i=1}^{t} \frac{i}{\alpha(i+1)} L^2 \le \frac{tL^2}{\alpha}$$

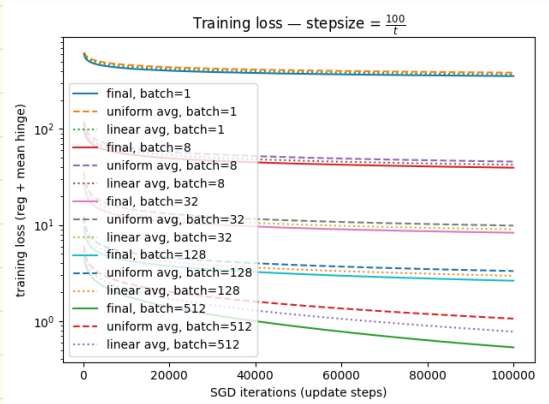
Divide by $\sum_{i=1}^{t} i = \frac{t(t+1)}{2}$ and continue ∎

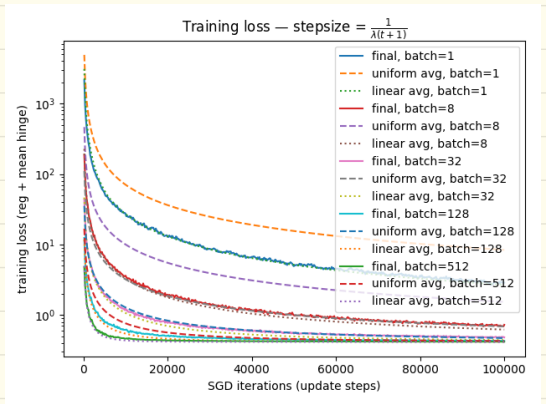
Regularized SVM

$$\min_x \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i x^T z_i\} + \frac{\lambda}{2}\|x\|^2$$

with data (z_i, y_i) with $z_i \in \mathbb{R}^{2000}$, $n = 10{,}000$, $\lambda = 10^{-3}$







Batch size does improve performance in experiments. In theory, this should not be the case because

$$\mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b} v_i\right\|_2^2 = \mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b}(v_i - \mathbb{E}v_i) + \frac{1}{b}\sum_{i=1}^{b}\mathbb{E}v_i\right\|^2$$

$$= \mathbb{E}\left\|\frac{1}{b}\sum_{i=1}^{b}(v_i - \mathbb{E}v_i)\right\|_2^2 + \left\|\frac{1}{b}\sum_{i=1}^{b}\mathbb{E}v_i\right\|^2$$

$\approx O(\frac{1}{b})$ $\quad$ $= \|\nabla f(x)\|_2^2$

with $v_i \in \partial f_i(x_i)$

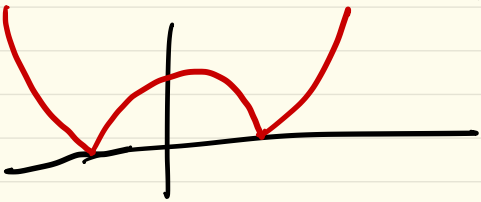
So

$$L^2 \simeq O\left(\frac{\sigma^2}{b}\right) + \|\nabla f(x)\|_2^2$$

If $\|\nabla f(x)\|$ is small (or gets smaller as you iterate) bigger batches help. This typical if $f(x) = \mathop{\mathbb{E}}_{z} \ell(x, z)$ is smooth.

If $\|\nabla f(x)\|$ is not small, then bigger batches don't help.

Eg: $f(x) = \mathop{\mathbb{E}}_{a \sim N(0, I)} |\langle x, a\rangle^2 - \langle x_\#, a\rangle^2|$



We now replace Lipschitz continuity by smoothness

Thm: Let $f:\mathbb{E}\to\mathbb{R}$ be β-smooth. and suppose that $g(x,z)$ satisfies

$$\mathop{\mathbb{E}}_{z} g(x,z) = \nabla f(x) \quad \text{and}$$

$$\mathop{\mathbb{E}}_{z} \|g(x,z)-\nabla f(x)\|_2^2 \le \sigma^2 \quad \forall x.$$

can be relaxed in many ways.

Then SGD with step-size $\eta_t \le \frac{1}{\beta}$ satisfies:

$$\mathbb{E}\|\nabla f(x_{i^*})\|^2 \le \frac{(f(x_0)-f^*) + (\sum_{i=1}^T \eta_i)\beta\sigma^2}{\sum_{i=1}^T \eta_i}$$

where $i^* \in \{1,\dots,T\}$ is chosen according to $P(i^*=i) = \eta_i / \sum_{j=1}^T \eta_j$.

pf: Compute for $v_t = g(x_t, z_t)$ the estimate

$$f(x_{t+1}) \le f(x_t) + \langle \nabla f(x_t), x_{t+1}-x_t\rangle + \frac{\beta}{2}\|x_{t+1}-x_t\|^2$$

$$= f(x_t) - \eta_t \langle \nabla f(x_t), v_t\rangle + \frac{\beta\eta_t^2}{2}\|v_t\|^2$$

Now take $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid x_t]$ to get

$$\mathbb{E}_t f(x_{t+1}) \le f(x_t) - \eta_t \|\nabla f(x_t)\|^2 + \frac{\eta_t^2 B}{2} \mathbb{E}\|v_t\|^2$$

Note

$$\mathbb{E}\|g(x,z)\|^2 = \underbrace{\mathbb{E}\|g(x,z) - \nabla f(x)\|^2}_{\le \sigma^2} + \|\nabla f(x)\|_2^2$$

So now:

$$\mathbb{E}_t f(x_{t+1}) \le f(x_t) - \eta_t \left(1 - \frac{\eta_t B}{2}\right) \|\nabla f(x_t)\|^2 + \frac{\eta_t^2 B \sigma^2}{2}$$

If $\eta_t \le \frac{1}{B}$, then $1 - \frac{\eta_t B}{2} \ge \frac{1}{2}$

So

$$\mathbb{E}_t f(x_{t+1}) \le f(x_t) - \frac{\eta_t}{2} \|\nabla f(x_t)\|^2 + \frac{\eta_t^2 \sigma^2 B}{2}.$$

Take $\mathbb{E}$ w.r.t. x_t to get

$$\mathbb{E} f(x_{t+1}) \le \mathbb{E} f(x_t) - \frac{\eta_t}{2} \mathbb{E}\|\nabla f(x_t)\|^2 + \frac{\eta_t^2 B\sigma^2}{2}$$

So iterate to get

$$\mathbb{E} f(x_{t+1}) \le f(x_0) - \sum_{i=1}^{t} \frac{\eta_i}{2} \mathbb{E}\|\nabla f(x_i)\|^2 + \left(\sum_{i=1}^{t} \eta_i^2\right) \cdot \frac{B\sigma^2}{2}$$

$\downarrow$ f^*

Rearrange:

$$\frac{\frac{1}{2}\sum_{i=1}^{t} \eta_i \mathbb{E}\|\nabla f(x_i)\|^2}{\sum_{i=1}^{t} \eta_i} \le \frac{\left(f(x_0) - f^* + \left(\sum_{i=1}^{t} \eta_i^2\right)\frac{B\sigma^2}{2}\right)}{\sum_{i=1}^{t} \eta_i}$$

Pick i^* with probability $P(i^* = i) = \frac{\eta_i}{\sum_{j=1}^{T} \eta_j}$

and returns $\|\nabla f(x_i)\|_2^2$. Then

$$\mathbb{E}\|\nabla f(x_{i^*})\|_2^2 = \sum_{i=1}^{T} \frac{\eta_i}{\sum_{j=1}^{T} \eta_j} \cdot \|\nabla f(x_i)\|^2$$

$$\leq \frac{f(x_0) - f^* + \left(\sum_{i=1}^{T} \eta_i^2\right) \beta \sigma^2}{\sum_{i=1}^{T} \eta_i}$$

In particular, if $\eta_t \equiv \eta < \frac{1}{\beta}$. Then

$$\boxed{\mathbb{E}\|\nabla f(x_{i^*})\|_2^2 \leq \frac{f(x_0) - f^*}{T\eta} + \eta \beta \sigma^2}$$

Thm: Suppose $f: \mathbb{E} \to \mathbb{R}$ is β-smooth and cvx. Suppose

$$\mathop{\mathbb{E}}_z g(x;z) = \nabla f(x), \quad \mathop{\mathbb{E}}_z \|g(x;z) - \nabla f(x)\|^2 \le \sigma^2 \quad \forall x.$$

Then for $\eta_t \equiv \eta \le \frac{1}{2\beta}$ get

$$\mathbb{E} f(\hat{x}_T) - f^* \le \frac{\|x_0 - \bar{x}\|^2}{2\eta T} + \frac{\eta}{2}\sigma^2.$$

where $\hat{x}_T = \frac{1}{T+1}\sum_{i=0}^{T} x_i$

pf: Compute for $g_t = g(x_t, z_t)$ to get

$$\|x_{t+1} - \bar{x}\|^2 = \|x_t - \eta g_t - \bar{x}\|^2$$
$$= \|x_t - \bar{x}\|^2 - 2\eta \langle g_t, x_t - \bar{x}\rangle + \eta^2 \|g_t\|^2$$

Take $\mathbb{E}_t$ to get

$$\mathbb{E}_t \|x_{t+1} - \bar{x}\|^2 = \|x_t - \bar{x}\|^2 - 2\eta \underbrace{\langle \nabla f(x_t), x_t - \bar{x}\rangle}_{\ge f(x_t) - f^*} + \eta^2 \mathbb{E}\|g_t\|^2$$

$$= \mathbb{E}\|g_t - \nabla f(x_t)\|^2 + \|\nabla f(x_t)\|^2$$

$$\mathbb{E}_t \|x_{t+1} - \bar{x}\|^2 \le \|x_t - \bar{x}\|^2 - 2\eta\,(f(x_t) - f^*) + \eta^2\sigma^2 + \eta^2 \cdot \|\nabla f(x_t)\|^2$$

Smoothness implies

(*) $\|\nabla f(x)\|^2 \le 2\beta\,(f(x) - f^*)$ we'll check this.

Then

$$\mathbb{E}_t \|x_{t+1} - \bar{x}\|^2 \le \|x_t - \bar{x}\|^2 - 2\eta\,\underbrace{(1 - \beta\eta)}_{\ge \frac{1}{2}}(f(x_t) - f^*) + \eta^2\sigma^2$$

Take $\mathbb{E}$ and rearrange:

$$\eta\,(\mathbb{E} f(x_t) - f^*) \le \mathbb{E}\|x_t - \bar{x}\|^2 - \mathbb{E}\|x_{t+1} - \bar{x}\|^2 + \eta^2\sigma^2$$

Sum up (right side telescopes), divide by $\eta(T+1)$ and lower bound left side by $f(\bar{x}_T) - f^*$ ☐

If you do step-decay get complexity

$$O\left(\beta\frac{\|x_0 - \bar{x}\|^2}{\varepsilon} + \frac{\sigma^2}{\varepsilon}\right)$$

Check of (*):

$$\underbrace{f(x) - f\left(x - \tfrac{1}{\beta}\nabla f(x)\right)}_{\| \\ f(x) - f^*} \geq \frac{1}{2\beta}\|\nabla f(x)\|^2$$

Consequences: If we optimize in η we get: $\eta = \frac{\|x_0 - \bar{x}\|}{\sigma\sqrt{T}}$

to get

$$\mathbb{E}\left(f(\bar{x}_T) - f^*\right) \leq O\left(\frac{\sigma \cdot \|x_0 - \bar{x}\|}{\sqrt{T}}\right)$$

So if $f(x) = \mathbb{E}_p\, \ell(x, z)$

and $g_t = \frac{1}{b}\sum_{i=1}^{b} \nabla \ell(x_t, z_i)$

for $z_1, \dots, z_b \overset{iid}{\sim} p$, then get $\frac{\sigma\|x_0 - \bar{x}\|}{b\cdot\sqrt{T}}$.

Thm: Suppose f is β-smooth and α-cvx. Suppose

$$\mathbb{E}\, g(x,z) = \nabla f(x), \qquad \mathbb{E}\|g(x,z) - \nabla f(x)\|^2 \le \sigma^2$$

Then with $\eta_t \le \frac{1}{4\beta}$ get

$$\eta_t \mathbb{E}(f(x_t) - f^*) \le (1 - \alpha\eta_t)\,\mathbb{E}\|x_t - \bar{x}\|^2 - \mathbb{E}\|x_{t+1} - \bar{x}\|^2 + \eta_t^2\sigma^2$$

pf: Compute for $g_t = g(x_t, z_t)$ get

$$\|x_{t+1} - \bar{x}\|^2 \le \|x_t - \bar{x} - \eta_t g_t\|^2$$

$$= \|x_t - \bar{x}\|^2 - 2\eta_t \langle g_t, x_t - \bar{x}\rangle + \eta_t^2 \|g_t\|^2$$

Take $\mathbb{E}_t$ and α-cvx to get

$$\mathbb{E}_t\|x_{t+1} - \bar{x}\|^2 \le \|x_t - \bar{x}\|^2 - 2\eta_t\left[f(x_t) - f^* + \frac{\alpha}{2}\|x_t - \bar{x}\|^2\right] + \eta_t^2\left[\|\nabla f(x)\|^2 + \sigma^2\right]$$

$$\le (1 - \alpha\eta_t)\|x_t - \bar{x}\|^2 - 2\eta_t(f(x_t) - f^*) + \left[2\beta\eta_t^2(f(x_t) - f^*) + \eta_t^2\sigma^2\right]$$

$$= (1-\alpha\eta_t)\|x_t-\bar{x}\|^2 - 2\eta_t(1-2\beta\eta_t)(f(x_t)-f^*) + \eta_t^2\sigma^2$$

$$\le (1-\alpha\eta_t)\|x_t-\bar{x}\|^2 - \eta_t(f(x_t)-f^*) + \eta_t^2\sigma^2$$

Take $\mathbb{E}$ and rearrange:

$$\eta_t \mathbb{E}(f(x_t)-f^*) \le (1-\alpha\eta_t)\mathbb{E}\|x_t-\bar{x}\|^2 - \mathbb{E}\|x_{t+1}-\bar{x}\|^2 + \eta_t^2\sigma^2$$

as claimed. ☐

So if $\eta_t \equiv \eta \le \frac{1}{4\beta}$, get

$$\eta\,\mathbb{E}(f(x_t)-f^*) \le (1-\alpha\eta)\mathbb{E}\|x_t-\bar{x}\|^2 - \mathbb{E}\|x_{t+1}-\bar{x}\|^2 + \eta^2\sigma^2$$

Define: $\Gamma_t = \prod_{i=0}^{t}(1-\alpha\eta)^i$. Divide by Γ_t:

$$\frac{\eta}{\Gamma_t}\mathbb{E}(f(x_t)-f^*) \le \frac{1}{\Gamma_{t-1}}\mathbb{E}\|x_t-\bar{x}\|^2 - \frac{1}{\Gamma_t}\mathbb{E}\|x_{t+1}-\bar{x}\|^2 + \frac{\eta^2\sigma^2}{\Gamma_t}$$

So sum up to get:

$$\sum_{i=1}^{t} \frac{\eta}{\Gamma_i} \mathbb{E}(f(x_i) - f^*) \leq \|x_0 - \bar{x}\|^2 + \sum_{i=0}^{t} \frac{\eta^2 \sigma^2}{\Gamma_i}$$

Algebraic Identity $\leq \frac{\eta \cdot \sigma^2}{2\Gamma_t}$

$$1 + \sum_{i=0}^{t} \frac{1}{\Gamma_i} = \frac{1}{\Gamma_t}$$

[Check!]

So the average iterate

$$\hat{x}_t = \frac{1}{\sum_{i=0}^{t} \frac{\eta}{\Gamma_i}} \sum_{i=0}^{t} \frac{\eta}{\Gamma_i} x_i$$

Satisfies

$$\mathbb{E} f(\hat{x}) - f^* \leq \frac{\Gamma_t}{\eta} \left(\|x_0 - \bar{x}\| + \eta^2 \sigma^2 \right)$$

Finally:

$$\mathbb{E}\, f(\hat{x}_t) - f^* \leq (1-\alpha\eta)^t \|x_0 - \bar{x}\|^2 + \frac{\eta\sigma^2}{\alpha}.$$

Now do step-decay to get complexity

$$\frac{B}{\alpha}\log\left(\frac{\|x_0-\bar{x}\|^2}{\varepsilon}\right) + \frac{\sigma^2}{\alpha\varepsilon}$$

Check!

Variance Reduction:

$$\min_x f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

where each f_i is β-smooth, f is α-cvx

Two options:

① Do GD $\Rightarrow$ complexity $O\left(\frac{\beta}{\alpha}\log\left(\frac{1}{\epsilon}\right)\right)$

② Do SGD $\Rightarrow$ complexity $O\left(\frac{\sigma^2}{\alpha\epsilon}\right)$ ← ignores variance

In terms of individual grad evals:

$$\underbrace{\frac{n\beta}{\alpha}\log\left(\frac{1}{\epsilon}\right)}_{GD} \quad \text{vs} \quad \underbrace{\frac{\sigma^2}{\alpha\epsilon}}_{SGD}$$

Typically $\alpha \approx \frac{1}{n}$. So $n^2\log\left(\frac{1}{\epsilon}\right)$ vs $\frac{n\sigma^2}{\epsilon}$ (GD) (SGD)

We'll get an algo with complexity

$$\left(n + \underbrace{\frac{\beta}{\alpha}}_{\approx n}\right)\log\left(\frac{1}{\epsilon}\right)$$

We'll talk about SVRG!

Idea is to reduce the variance of ∇f_i where i is sample uniformly from $\{1,\dots,n\}$.

Candidate: Given $x_1,\dots,x_t$ look at

$$\nabla f_i(x_t) - \nabla f_i(y) + \nabla f(y)$$

(y: anchor point)

Remark: y should ideally satisfy $y \approx \bar{x}$.

Lemma:

$$\mathbb{E}\,\|\nabla f_i(x) - \nabla f_i(\bar{x})\|_2^2 \le 2\beta(f(x) - f^*)$$

pf: Define $h(x) = f_i(x) - f_i(\bar{x}) - \langle \nabla f_i(\bar{x}), x - \bar{x}\rangle$

Then since h is β-smooth, get:

$$\|\nabla h(x)\|^2 \le 2\beta\,(h(x) - \underbrace{\min h}_{=0 \text{ // cvx}})$$

So

$$\|\nabla f_i(x) - \nabla f_i(\bar{x})\|^2 \le 2\beta h(x)$$

Note $\mathbb{E}\,h(x) = 2\beta(f(x) - f^*)$ ∎

Consequently

$$\mathbb{E}\|\nabla f_i(x) - \nabla f_i(y) + \nabla f(y)\|^2$$

$$= \mathbb{E}\|\nabla f_i(x) - \nabla f_i(\bar{x}) + \nabla f_i(\bar{x}) - \nabla f_i(y) + \nabla f(y)\|^2$$

$$\left[(a+b)^2 \le 2a^2 + 2b^2\right]$$

$$\le 2\underbrace{\mathbb{E}\|\nabla f_i(x) - \nabla f_i(\bar{x})\|^2}_{\text{Lemma}} + 2\underbrace{\mathbb{E}\|[\nabla f_i(\bar{x}) - \nabla f_i(y)] + \nabla f(y)\|^2}_{\mathbb{E}\|\nabla f_i(\bar{x}) - \nabla f_i(y)\|^2 \le 2B(f(y) - f^*) \text{ (Lemma)}}$$

$2B(f(x) - f^*)$

(*)

SVRG: Let $y^{(1)} \in \mathbb{E}$. For $s = 1, 2, \ldots$, let $x_1^{(s)} = y^{(s)}$. For $t = 1, \ldots, k$ let

$$x_{t+1}^{(s)} = x_t^{(s)} - \eta\left(\nabla f_{i_t^{(s)}}(x_t^{(s)}) - \nabla f_{i_t^{(s)}}(y^{(s)}) + \nabla f(y^{(s)})\right)$$

where $i_t^{(s)}$ is iid $\sim \{1, \ldots, n\}$. Let

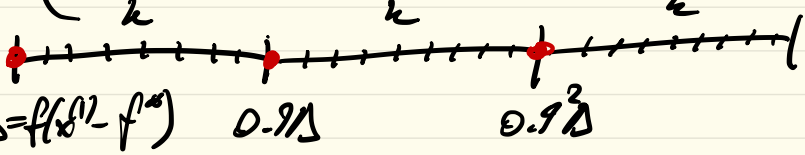
$$y^{(s+1)} = \frac{1}{k}\sum_{t=1}^{k} x_t^{(s)}$$

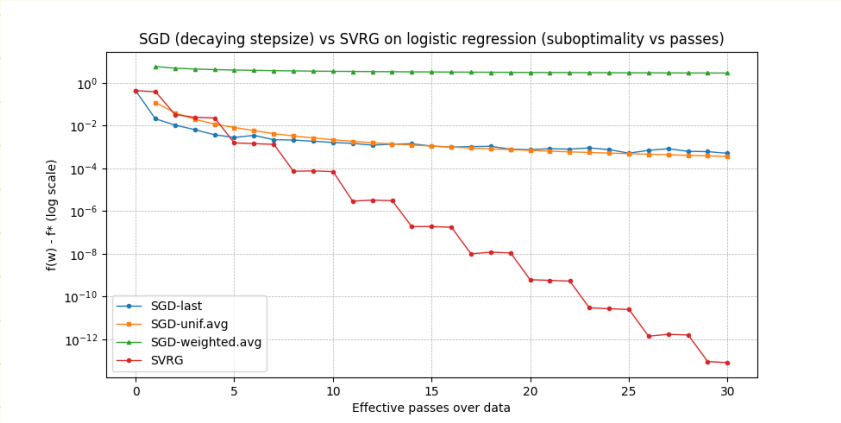
Thm: SVRG with $\eta = \frac{1}{10B}$ and $k = 20\frac{B}{\alpha}$ satisfies

$$\mathbb{E} f(y^{(s+1)}) - f^* \leq 0.9^s (f(y^{(1)}) - f^*)$$

So after $\frac{1}{0.9} \log\left(\frac{f(x^{(1)}) - f^*}{\epsilon}\right)$ epochs we have $\mathbb{E} f(y^{(s+1)}) - f^* \leq \epsilon$. Each epoch uses $O(\frac{B}{\alpha})$ individual gradients. So total # ind. grad evals is

$$O\left(\left(n + \frac{B}{\alpha}\right) \log\left(\frac{f(x^{(1)}) - f^*}{\epsilon}\right)\right).$$





pf: Let's look at a single epoch initialized at anchor y.

Set $v_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(y) + \nabla f(y)$

Then

$$\|x_{t+1} - \bar{x}\|^2 = \|x_t - \bar{x}\|^2 - 2\eta\langle v_t, x_t - \bar{x}\rangle + \eta^2\|v_t\|^2$$

Take $\mathbb{E}_t$ to get

$$\mathbb{E}_t\|x_{t+1} - \bar{x}\|^2 = \|x_t - \bar{x}\|^2 - 2\eta\underbrace{\langle \nabla f(x_t), x_t - \bar{x}\rangle}_{\geq f(x_t) - f^*} + \eta^2\mathbb{E}_t\|v_t\|^2$$

$$\mathbb{E}_t\|v_t\|^2 \leq 2\beta(f(x_t) - f^*) + 2\beta(f(y) - f^*)$$

Combining

$$\leq \|x_t - \bar{x}\|^2 - 2\eta(1 - \beta\eta)(f(x_t) - f^*) + 2\beta\eta^2(f(y) - f^*)$$

Take $\mathbb{E}$ and iterate:

$$\mathbb{E}\|x_{t+1} - \bar{x}\|^2 \le \|y - \bar{x}\|^2 - 2\eta(1-\eta\beta)\,\mathbb{E}\sum_{i=1}^{k}(f(x_i) - f^*) + 2B\eta^2 k\,(f(y) - f^*)$$

Recall: $\frac{\alpha}{2}\|y - \bar{x}\|^2 \le (f(y) - f^*)$ (Cvx)

So rearrange

$$\mathbb{E}\,\frac{1}{k}\sum_{i=1}^{k}(f(x_i) - f^*) \le \left(\frac{2}{\alpha} + 2B\eta^2 k\right)\frac{(f(y) - f^*)}{2\eta(1-\eta\beta)\cdot k}$$

(Cvx) $\mathbb{E}\left(f\left(\frac{1}{k}\sum_{i=1}^{k} x_i\right) - f^*\right)$

$\Sigma_{\text{next anchor}}$

$$= \left(\frac{1}{\eta(1-\eta B)k\alpha} + \frac{B\eta}{1-\eta\beta}\right)(f(y) - f^*)$$

≤ 0.9

Done! ∎

Note for some problems the variance actually shrinks to zero automatically without any variance reduction.

Recall:

$$\mathbb{E}\|\nabla f_i(x) - \nabla f_i(\bar{x})\|_2^2 \le 2\beta(f(x) - f^*)$$

What if $\nabla f_i(\bar{x}) = 0 \quad \forall i$

↑ means $\bar{x}$ is a minimizer of $f_i \ \forall i$

Problems satisfying such a condition are called interpolation problems