

ML and Optimization Lecture 13

- Newton's Method: uses curvature information, so needs fewer iterations to reach high-quality solutions; each iteration is expensive
- Stochastic Gradient Descent: touches fewer data points, so each iteration is faster, but convergence to high-quality solutions is slower.

Newton's Method

Gradient Descent Update:

$$\begin{aligned}x_{t+1} &= \underset{x}{\operatorname{argmin}} f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \underbrace{\frac{1}{2\alpha_t} \|x - x_t\|_2^2}_{\text{we model the curvature of } f \text{ locally at } x_t \text{ using } \nabla^2 f(x_t) \approx \frac{1}{\alpha_t} I}} \\ &= x_t - \alpha_t \nabla f(x_t)\end{aligned}$$

Newton's method: use the optimal local quadratic approximation

$$\begin{aligned}x_{t+1} &= \underset{x}{\operatorname{argmin}} \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2} (x - x_t)^T \nabla^2 f(x_t) (x - x_t)}_{\text{we model the curvature of } f \text{ locally at } x_t \text{ using } \nabla^2 f(x_t) \approx \frac{1}{\alpha_t} I}} \\ &= x_t - \nabla^2 f(x_t)^{-1} \nabla f(x_t) \quad := m(x)\end{aligned}$$

because x_* satisfies $\nabla_x m(x^*) = 0 = \nabla f(x_t) + \nabla^2 f(x_t)(x - x_t)$

"Unguarded Newton method/update"

We generally take our search direction

$$g_t = -[\nabla^2 f(x_t)]^{-1} \nabla f(x_t)$$

and judiciously choose our step-size α_t to ensure $x_{t+1} = x_t + \alpha_t g_t$

$$f(x_{t+1}) = f(x_t + \alpha_t g_t) < f(x_t)$$

e.g. with backtracking line search. This is called the **guarded Newton's method**.

There are two phases to this algorithm:

- "**damped phase**" (when $\alpha_t < 1$): suboptimality reduces linearly:

$$f(x_t) - f(x_*) \leq \varepsilon \text{ after } t = O(\log(\frac{1}{\varepsilon}))$$

- "**purely Newton phase**" (when $\alpha_t = 1$): quadratic convergence at each iteration, the number of digits of precision doubles!

$$f(x_t) - f(x_*) \leq \varepsilon \text{ after } t = O(\log \log \frac{1}{\varepsilon})$$

Ex. Consider an ill-conditioned linear regression problem

$$x^* = \operatorname{argmin}_x \frac{1}{2} \|Ax - b\|_2^2$$

$$\Rightarrow \nabla^2 f(x) = A^T A$$

and $\lambda_{\min}(A^T A) \cdot I \preceq A^T A \preceq \lambda_{\max}(A^T A) \cdot I$

so

$$\kappa_f = \frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)}$$

assume

$$\lambda_{\max}(A^T A) \gg 1, \quad \lambda_{\min}(A^T A) \ll 1$$

and

$$\kappa_f = e^C \quad \text{where } C > 0 \text{ is large}$$

Option 1

Use GD with stepsize $\frac{1}{\beta}$ where $\beta = \lambda_{\max}(A^T A) \gg 1$

$$f(x_T) - f(x_*) \leq \varepsilon \quad \text{when} \quad T = O\left(\kappa_f \log\left(\frac{1}{\varepsilon}\right)\right) \gg 1$$

Option 2

Use Newton's method with unit step size

$$x_1 = x_0 - (A^T A)^{-1} A^T (A x_0 - b)$$

$$= x_0 - (x_0 - (A^T A)^{-1} A^T b)$$

$$= (A^T A)^{-1} A^T b$$

$$= x_{\text{OLS}}$$

$$= \operatorname{argmin}_x \frac{1}{2} \|Ax - b\|_2^2$$

one iteration of
Newton's method
gives the exact
solution

Takeaway: we converge
faster if we account for
curvature

Drawbacks to Newton's Method

- need to form $\nabla^2 f(x_t) \in \mathbb{R}^{d \times d}$
and solve a linear system

In general this costs $O(d^2)$ memory and $O(d^3)$ operations to compute the Newton search direction
so when d is large, Newton's method is infeasible

GD can be expensive per iteration

Ex: GD to solve an ERM problem

$$f(w) = \frac{1}{n} \sum_{i=1}^n \ell_i(w)$$

$\ell_i(w)$ is loss on
on i th datapoint, e.g.

$$\ell_i(w) = \log(1 + \exp(-y_i w^T x_i))$$

or

$$\ell_i(w) = (w^T x_i - y_i)^2$$

$$\Rightarrow \nabla f(w) = \frac{1}{n} \sum_{i=1}^n \nabla \ell_i(w)$$

Per iteration to compute the full gradient

$\nabla f(w)$ requires touching $n \gg 1$ datapoints!

$\mathcal{O}(nd)$ operations

Stochastic Gradient Descent

We want an algorithm that's cheaper per iteration than GD, preferably $\mathcal{O}(d)$ per iteration

Idea: notice that $\nabla f(\omega_t)$ is an approximate direction to search for decrease

$$f(\omega) \approx f(\omega_t) + \langle \nabla f(\omega_t), \omega - \omega_t \rangle + \frac{1}{2\alpha_t} \|\omega - \omega_t\|_2^2$$

where

$$f(\omega) = \frac{1}{n} \sum_{i=1}^n \ell_i(\omega)$$

$\Rightarrow$

$$\nabla f(\omega) = \frac{1}{n} \sum_{i=1}^n \nabla \ell_i(\omega)$$

$\nabla f(\omega)$ is already a sample average that approximates a population average

Treat $\nabla f(\omega)$ itself as a population average, and approximate it with a sample average

- select a random subset $I \subset [n] = \{1, \dots, n\}$ of size k and form an approximation

$$g = \frac{1}{k} \sum_{i \in I} \nabla \ell_i(\omega) \quad \text{--- stochastic gradient estimate}$$

if we chose I appropriately, then

$$\mathbb{E} g = \nabla f(\omega) \quad \text{--- the stochastic gradient estimate is an unbiased estimator of the gradient}$$

The set I is called a minibatch

Minibatch Stochastic Gradient Descent

$$x_* \in \operatorname{argmin}_x f(x)$$

Choose a minibatch size k

for $t=0, \dots, T-1$

- sample $I \subseteq [n]$ of size k (independently between iterations)

$$- g_t \leftarrow \frac{1}{k} \sum_{i \in I} \nabla \ell_i(x_t)$$

$$- x_{t+1} = x_t - \alpha_t g_t$$

return:

$$- x_T$$

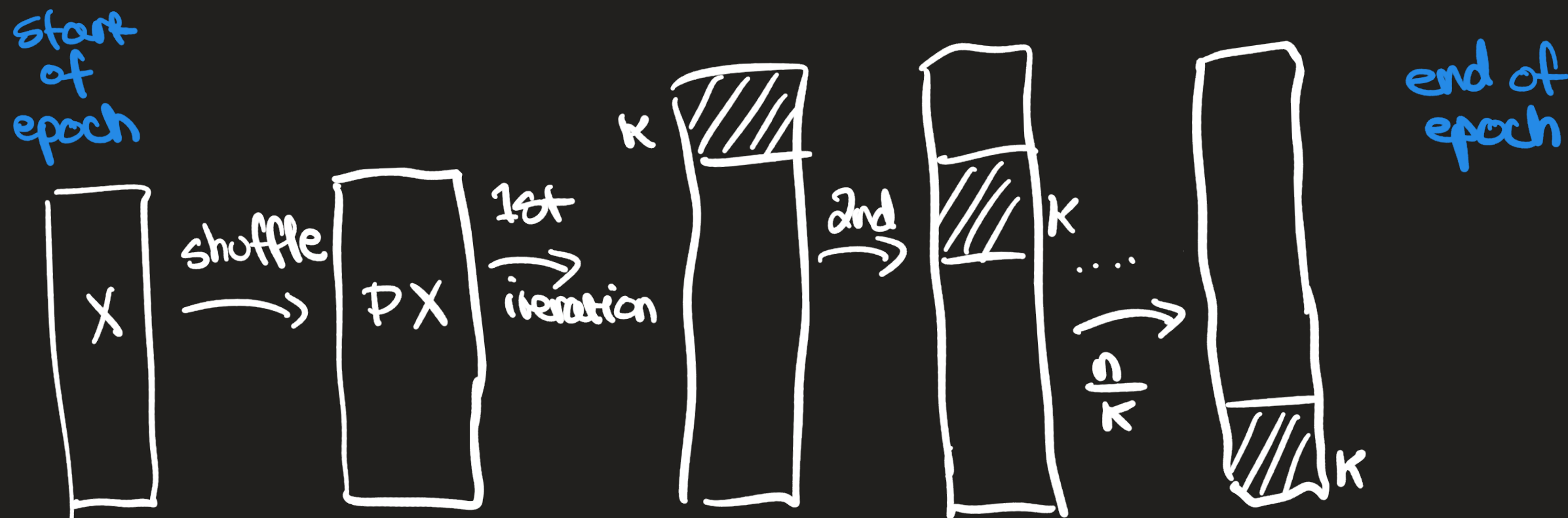
or

$$- \bar{x} = \frac{1}{T} \sum_{i=1}^T x_i$$

In practice, to efficiently use our training data, we use Randomized Reshuffling:

In each epoch:

- randomly shuffle the training data
- everytime we sample a minibatch, use the first k unused points



notice: there are $\frac{N}{K}$ minibatches per epoch

RR Minibatch SGD

Inputs: ∇f oracle, α_t stepsizes, K minibatch size,
 E epochs

Alg

$$z_0 = x_0$$

for $e=1, \dots, E$

$(x, y) \leftarrow$ random reshuffle

$$x_0 = z_{e-1}$$

for $t=0, \dots, \frac{n}{K} - 1$

$I \leftarrow$ first K unused samples in this epoch

$$g_t \leftarrow \frac{1}{K} \sum_{i \in I} \nabla \ell_i(x_t)$$

$$x_{t+1} \leftarrow x_t - \alpha g_t$$

end

$$z_e = x_{\frac{n}{K}}$$

end

return z_E

How does Minibatch SGD perform?

- We don't care about κ , per se. Instead we assume

$$\mathbb{E}[\|g_t - \nabla f(x_t)\|^2] \leq \sigma^2$$

or $\mathbb{E}\|g_t\|^2 \leq \sigma^2$

called the noise level of the gradient estimate as $\kappa \uparrow$, $\sigma^2 \downarrow$

- We always assume $\mathbb{E}[g_t | x_t] = \nabla f(x_t)$

- We will look at performance for μ -strongly convex objectives to understand impact of noise level σ^2

- We will look at convergence guarantees for nonconvex optimization as well

Monotonicity

If f is a μ -strongly convex function then

$$\forall x, y \in \text{dom}(f): \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$$

i.e. the gradient operator is monotone

Prf

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_2^2$$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2$$

$$\Rightarrow f(x) + f(y) \geq f(y) + f(x) + \langle \nabla f(y) - \nabla f(x), x - y \rangle + \mu \|x - y\|_2^2$$

$$\Rightarrow \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|_2^2$$



Consider f to be μ -strongly convex to see the impact of the noisy gradient estimate on the convergence of x_t

If $\alpha < \frac{1}{2\mu}$ and $\mathbb{E} \|g\|_2^2 \leq \sigma^2$, then

$$\mathbb{E} \|x_T - x_*\|_2^2 \leq (1 - 2\alpha\mu)^T \|x_0 - x_*\|_2^2 + \frac{\alpha\sigma^2}{2\mu}$$

so SGD converges to within a radius of the minimizer x_* that is determined by the noise level σ^2



Takeaway: to converge to x^* , unlike in GD, you must use a shrinking stepsize

Pf

$$\begin{aligned}x_{t+1} - x_* &= x_t - \alpha g_t - x_* \\ &= x_t - x_* - \alpha g_t\end{aligned}$$

so

$$\begin{aligned}\mathbb{E} \|x_{t+1} - x_*\|_2^2 &= \mathbb{E} \|x_t - x_* - \alpha g_t\|_2^2 \\ &= \mathbb{E} \left[\|x_t - x_*\|_2^2 - 2 \langle x_t - x_*, \alpha g_t \rangle + \|\alpha g_t\|_2^2 \right] \\ &= \mathbb{E} \|x_t - x_*\|_2^2 + \alpha^2 \sigma^2 - 2 \mathbb{E} [\langle x_t - x_*, \alpha g_t \rangle] \\ &= \mathbb{E} \|x_t - x_*\|_2^2 + \alpha^2 \sigma^2 - 2\alpha \mathbb{E} [\langle x_t - x_*, \nabla f(x_t) \rangle] \\ &= \mathbb{E} \|x_t - x_*\|_2^2 + \alpha^2 \sigma^2 - 2\alpha \mathbb{E} \langle x_t - x_*, \nabla f(x_t) - \nabla f(x_*) \rangle\end{aligned}$$

Recall by monotonicity,

$$\langle x_t - x^*, \nabla f(x_t) - \nabla f(x^*) \rangle \geq \mu \|x_t - x^*\|_2^2$$

so

$$\begin{aligned} \mathbb{E} \|x_{t+1} - x^*\|_2^2 &\leq \mathbb{E} \|x_t - x^*\|_2^2 + \sigma^2 \alpha^2 - 2\alpha\mu \mathbb{E} \|x_t - x^*\|_2^2 \\ &= (1 - 2\alpha\mu) \mathbb{E} \|x_t - x^*\|_2^2 + \alpha^2 \sigma^2 \end{aligned}$$

To solve this recursion, let

$$\Delta_t = \mathbb{E} \|x_t - x^*\|_2^2 \quad \text{and} \quad c = (1 - 2\alpha\mu) \quad (\text{note } c < 1)$$
$$b = \alpha^2 \sigma^2$$

then

$$\Delta_{t+1} \leq c \Delta_t + b$$

$$\Rightarrow \Delta_1 \leq c \Delta_0 + b$$

$$\Delta_2 \leq c \Delta_1 + b \leq c^2 \Delta_0 + cb + b$$

$$\Delta_3 \leq c \Delta_2 + b \leq c^3 \Delta_0 + c^2 b + cb + b$$

In general,

$$\Delta_T \leq C^T \Delta_0 + \left(\sum_{i=0}^{T-1} C^i \right) b$$

$$\leq C^T \Delta_0 + \frac{1}{1-C} b$$

$$= C^T \Delta_0 + \frac{1}{2\alpha\mu} \alpha^2 \sigma^2$$

$$= C^T \Delta_0 + \frac{\alpha \sigma^2}{2\mu}$$

or

$$\mathbb{E} \|x_T - x^*\|_2^2 \leq (1 - 2\alpha\mu)^T \mathbb{E} \|x_0 - x^*\|_2^2 + \frac{\alpha \sigma^2}{2\mu}$$



Proof of convergence in pure Newton phase (not covered in class)

Newton's method locally converges quadratically fast, even for nonconvex objectives

Assume the Hessian of f is L -Lipschitz continuous,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L \|x - y\|_2$$

and strictly positive at x^* where $\nabla f(x^*) = 0$,

$$\nabla^2 f(x^*) \succcurlyeq \mu I$$

for some $\mu \geq 0$.

Then if the first iterate is within the basin of attraction,

$$\|x_0 - x^*\|_2 \leq \frac{\mu}{2L},$$

Newton's method is well-defined (i.e. $\nabla^2 f(x_t)$ is invertible), the iterates remain within the basin of attraction, and they converge to x^* at a quadratic rate:

$$\|x_{t+1} - x^*\|_2 \leq \frac{1}{\mu} \|\nabla f(x_t)\|_2^2$$

Prf

It suffices to show that if x_t is in the basin of attraction then $\|x_{t+1} - x^*\|_2 \leq \frac{L}{\mu} \|x_t - x^*\|_2^2$, (1) because this implies

$$\begin{aligned} \|x_{t+1} - x^*\|_2 &\leq \frac{L}{\mu} \cdot \frac{\mu}{2L} \|x_t - x^*\|_2 \leq \frac{1}{2} \|x_t - x^*\|_2 \\ &\leq \frac{\mu}{2L} \end{aligned}$$

so x_{t+1} is again in the basin of attraction.

To show (1), we compute

$$\begin{aligned} x_{t+1} - x^* &= x_t - \nabla^2 f(x_t)^{-1} \nabla f(x_t) - x^* \\ &= x_t - \nabla^2 f(x_t)^{-1} [\nabla f(x_t) - \nabla f(x^*)] - x^* \\ &= \nabla^2 f(x_t)^{-1} [\nabla f(x^*) - \nabla f(x_t) - \nabla^2 f(x_t)(x^* - x_t)] \end{aligned}$$

Now we expect two things intuitively:

1) because x_t is close to x^* and the Hessian is Lipschitz, $\nabla^2 f(x_t)$ should also be invertible (so the method is well-defined)

2) by a T.S. argument, because x_t is close to x^* ,

$$\nabla f(x^*) - \nabla f(x_t) - \nabla^2 f(x_t)(x^* - x_t) = O(\|x_t - x^*\|_2^2)$$

together these argue that we can expect

$$\|x_{t+1} - x^*\|_2 \leq O(\|x_t - x^*\|_2^2)$$

First, let's quantify the first point. In general, if A, B are matrices,

$$-\|A - B\|_2 \cdot I \preceq A - B \preceq \|A - B\|_2 \cdot I,$$

which implies

$$A \succeq B - \|A - B\|_2 \cdot I$$

Apply this fact to see that

$$\nabla^2 f(x_t) \succeq \nabla^2 f(x^*) - \|\nabla^2 f(x_t) - \nabla^2 f(x^*)\|_2 \cdot I$$

$$\succeq \mu \cdot I - L \|x_t - x^*\|_2 \cdot I$$

$$\succeq \mu \cdot I - L \cdot \frac{\mu}{2L} \cdot I$$

$$= \frac{\mu}{2} \cdot I$$

so indeed $\nabla^2 f(x_t)$ is invertible, and $\nabla^2 f(x_t)^{-1} \preceq \frac{2}{\mu} \cdot I$.

Next, we quantify the second claim with a T.S. argument.

Define

$$g(u) = \nabla f(x_t + u(x^* - x_t)) \quad , \quad \text{so}$$

$$g(1) - g(0) = \nabla f(x^*) - \nabla f(x_t) = \int_0^1 \nabla^2 f(x_t + u(x^* - x_t)) \cdot (x^* - x_t) du$$

so

$$\nabla f(x^*) - \nabla f(x_t) - \nabla^2 f(x_t)(x^* - x_t) = \int_0^1 [\nabla^2 f(x_t + \nu(x^* - x_t)) - \nabla^2 f(x_t)] (x^* - x_t) d\nu,$$

so

$$\begin{aligned} \|\nabla f(x^*) - \nabla f(x_t) - \nabla^2 f(x_t)(x^* - x_t)\|_2 &\leq \\ &\int_0^1 L \|\nu(x^* - x_t)\|_2 \|x^* - x_t\|_2 d\nu \\ &= \frac{L}{2} \|x^* - x_t\|_2^2 \end{aligned}$$

Putting the pieces together,

$$x_{t+1} - x_* = \nabla f(x_t)^{-1} [\nabla f(x_*) - \nabla f(x_t) - \nabla^2 f(x_t)(x_* - x_t)]$$

so

$$\|x_{t+1} - x_*\|_2 \leq \|\nabla f(x_t)^{-1}\|_2 \cdot$$

$$\|\nabla f(x_*) - \nabla f(x_t) - \nabla^2 f(x_t)(x_* - x_t)\|_2$$

$$\leq \frac{2}{\mu} \cdot \frac{L}{2} \|x_t - x_*\|_2^2$$

$$= \frac{L}{\mu} \|x_t - x_*\|_2^2,$$

as claimed.

