ML and Optimization Lecture 12

- 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

$$\nabla f(x_t) + \frac{1}{\alpha_t}(x_{t+1} - x_t) = 0$$

Gradient Descent Update:

$$x_{t+1} = \underset{x}{\operatorname{argmin}}\ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \underbrace{\frac{1}{2\alpha_t}\|x - x_t\|_2^2}$$

$$= x_t - \alpha_t \nabla f(x_t)$$

we model the curvature of f locally at x_t using $\nabla^2 f(x_t) \approx \frac{1}{\alpha_t} I$

Newton's method: use the optimal local quadratic approximation

$$x_{t+1} = \underset{x}{\operatorname{argmin}}\ \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2}(x - x_t)^T \nabla^2 f(x_t)(x - x_t)}_{:= m(x)}$$

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

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

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

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:

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

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

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

Ex. Consider an ill-conditioned linear regression problem

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

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

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

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

assume $$\lambda_{max}(A^TA) \gg 1 \;,\; \lambda_{min}(A^TA) \ll 1$$

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

Option 1

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

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

Option 2

Use Newton's method with unit step size

$$\begin{aligned} x_1 &= x_0 - (A^TA)^{-1}A^T(Ax_0 - b) \\ &= x_0 - (x_0 - (A^TA)^{-1}A^Tb) \\ &= (A^TA)^{-1}A^Tb \\ &= x_{ols} \\ &= \operatorname{argmin}_x \frac{1}{2}\|Ax-b\|_2^2 \end{aligned}$$

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

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

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

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

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 $n >> 1$ datapoints!

$O(nd)$ operations

Stochastic Gradient Descent

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

Idea: notice that $\nabla f(w_t)$ is an <u>approximate</u> direction to search for decrease

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

where

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

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

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

Treat $\nabla f(w)$ 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

$$g = \frac{1}{k} \sum_{i \in I} \nabla \ell_i(w)$$ — stochastic gradient estimate

if we chose I appropriately, then

$$\mathbb{E}\, g = \nabla f(w)$$ — 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)
- $g_t \leftarrow \frac{1}{k} \sum_{i \in I} \nabla l_i(x_t)$
- $x_{t+1} = x_t - \alpha_t g_t$

return:

- x_T

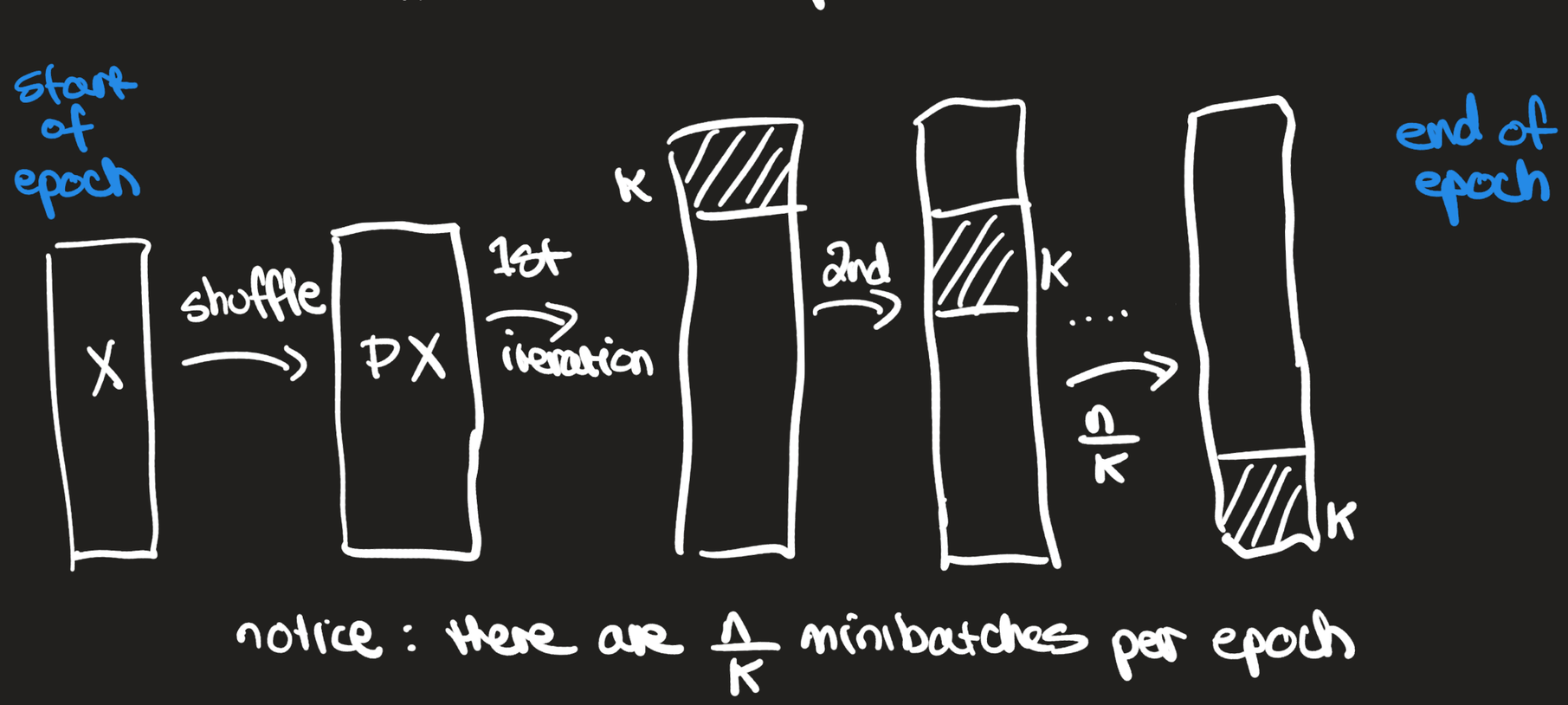
or

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

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 K_γ per se. Instead we assume

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

↪ called the noise level of the gradient estimate

as $K \uparrow$, $\sigma^2 \downarrow$

or

$$\mathbb{E}\|g_t\|_2^2 \leq \sigma^2$$

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

- We will look at performance for μ-strongly cvx 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

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

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

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

∎

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

If $\alpha \leq \frac{1}{2\mu}$ 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

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

$$= x_t - x_* - \alpha g_t$$

so

$$\mathbb{E}\|x_{t+1} - x_*\|_2^2 = \mathbb{E}\|x_t - x_* - \alpha g_t\|_2^2$$

$$= \mathbb{E}\left[\|x_t - x_*\|_2^2 - 2\langle x_t - x_*, \alpha g_t\rangle + \|\alpha g_t\|_2^2\right]$$

$$= \mathbb{E}\|x_t - x_*\|_2^2 + \alpha^2\sigma^2 - 2\,\mathbb{E}\left[\langle x_t - x_*, \alpha g_t\rangle\right]$$

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

$$= \mathbb{E}\|x_t - x_*\|_2^2 + \alpha^2\sigma^2 - 2\alpha\,\mathbb{E}\langle x_t - x_*, \nabla f(x_t) - \nabla f(x^*)\rangle$$

Recall by monotonicity)

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

so

$$\mathbb{E}\|x_{t+1} - x^*\|_2^2 \leq \mathbb{E}\|x_t - x^*\|_2^2 + \sigma^2\alpha^2 - 2\alpha\mu\,\mathbb{E}\|x_t - x^*\|_2^2$$

$$= (1 - 2\alpha\mu)\,\mathbb{E}\|x_t - x^*\|_2^2 + \alpha^2\sigma^2$$

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

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

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

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

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

In general,

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

$$\le c^T \Delta_0 + \frac{1}{1-c} b$$

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

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

or

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

Convergence rate of SGD (for nonconvex opt prob)

Due to nonconvexity and fact SGD is not a descent method, we measure convergence with

$$\min_{t=0,\dots,T-1} \mathbb{E}\,\|\nabla f(x_t)\|_2^2 \leq \varepsilon$$

Assumptions

- f bounded below
- f is β-smooth
- assume $\mathbb{E}\, g_t = \nabla f(x_t)$
- $\mathbb{E}\,\|g_t\|_2^2 \leq \sigma^2$ (could relax to $\mathbb{E}\,\|\nabla f(x_t) - g_t\|_2^2 \leq \sigma^2$)

Idea: bound

$$\sum_{t=0}^{T-1} \alpha_t \mathbb{E} \|\nabla f(x_t)\|_2^2 \leq B$$

which implies

$$\min_{t=0,\cdots,T-1} \mathbb{E} \|\nabla f(x_t)\|_2^2 \leq \frac{B}{\sum_{t=0}^{T-1} \alpha_t}$$

Prf

Recall that b/c f is β-smooth,

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

$\Rightarrow$

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), -\alpha_t g_t \rangle + \frac{\beta}{2} \|-\alpha_t g_t\|_2^2$$

$$= f(x_t) - \alpha_t \langle \nabla f(x_t), g_t \rangle + \frac{\alpha_t^2 \beta}{2} \|g_t\|_2^2$$

Take expectations (b/c $x_t > x_{t+1} > g_t$ are random vectors)

$$\mathbb{E} f(x_{t+1}) \leq \mathbb{E} f(x_t) - \alpha_t \mathbb{E} \|\nabla f(x_t)\|_2^2 + \frac{\alpha_t^2 \beta \sigma^2}{2}$$

this implies

$$\alpha_t \mathbb{E} \|\nabla f(x_t)\|_2^2 \leq \mathbb{E}\left[f(x_t) - f(x_{t+1})\right] + \frac{\alpha_t^2 \beta \sigma^2}{2}$$

so

$$\sum_{t=0}^{T-1} \alpha_t \mathbb{E} \|\nabla f(x_t)\|_2^2 \leq \sum_{t=0}^{T-1} \mathbb{E}\left[f(x_t) - f(x_{t+1})\right]$$

↑ telescoping sum

$$+ \frac{\beta \sigma^2}{2} \sum_{t=0}^{T-1} \alpha_t^2$$

$$= f(x_0) - \mathbb{E} f(x_T) + \frac{\beta \sigma^2}{2} \sum_{t=0}^{T-1} \alpha_t^2$$

$$\leq f(x_0) + \frac{\beta \sigma^2}{2} \sum_{t=0}^{T-1} \alpha_t^2$$

So finally

$$\underbrace{\min_{t=0,\dots,T-1} \mathbb{E}\,\|\nabla f(x_t)\|_2^2}_{:= \text{gap}(T)} \leq \frac{f(x_0)}{\sum_{t=0}^{T-1}\alpha_t} + \frac{B\sigma^2}{2}\frac{\sum_{t=0}^{T-1}\alpha_t^2}{\sum_{t=0}^{T-1}\alpha_t}$$

Choices of step-sizes:

- $\alpha_t = \alpha$, then $\text{gap}(T) = O\left(\frac{f(x_0)}{T\alpha} + \frac{B\sigma^2}{2}\left(\frac{T\alpha^2}{T\alpha}\right)\right)$

$$= O\left(\frac{f(x_0)}{T} + \sigma^2\alpha\right)$$

- $\alpha_t = \frac{\alpha}{t+1}$, then $\sum_{t=0}^{T-1}\alpha_t = \sum_{t=0}^{T-1}\frac{\alpha}{t+1} = O(\ln T)$

and $\sum_{t=0}^{T-1}\alpha_t^2 = \sum_{t=0}^{T-1}\frac{\alpha^2}{(t+1)^2} = O(1) \Rightarrow \text{gap}(T) = O\left(\frac{1}{\ln T}\right)$

$$-\alpha_t = \frac{\alpha}{\sqrt{t+1}} \quad \text{then} \quad \sum_{t=0}^{T-1} \alpha_t = \sum_{t=0}^{T-1} \frac{\alpha}{\sqrt{t+1}} = O(\sqrt{T})$$

$$\text{and} \quad \sum_{t=0}^{T-1} \alpha_t^2 = O(\ln T), \quad \text{so}$$

$$\text{gap}(T) = O\left(\frac{\ln T}{\sqrt{T}}\right)$$

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:

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

Pf

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

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

$$\leq \frac{\mu}{2L}$$

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

To show (1), we compute

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

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

$$= \nabla^2 f(x_t)^{-1}\left[\nabla f(x_*) - \nabla f(x_t) - \nabla^2 f(x_t)(x^* - x_t)\right]$$

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

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

Apply this fact to see that

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

$$\succcurlyeq \mu \cdot I - L\|x_t - x_*\|_2 \cdot I$$

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

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

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

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

Define

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

$$g(1) - g(0) = \nabla f(x^*) - \nabla f(x_t) = \int_0^1 \nabla^2 f(x_t + v(x^* - x_t)) \cdot (x^* - x_t)\, dv$$

so

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

$$\int_0^1 \left[\nabla^2 f(x_t + v(x^* - x_t)) - \nabla^2 f(x_t)\right](x^* - x_t)\, dv,$$

so

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

$$\int_0^1 L \|v(x^* - x_t)\|_2 \|x^* - x_t\|_2 \, dv$$

$$= \frac{L}{2} \|x^* - x_t\|_2^2$$

Putting the pieces together,

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

so

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

$$\le \frac{2}{\mu} \cdot \frac{L}{2} \|x_t - x^*\|_2^2$$

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

as claimed.