Nonsmooth optimization

Bernd Gärtner, Martin Jaggi

1 Subgradient and subdifferentiation

Definition 1.1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. A vector $\mathbf{g}$ is a subgradient of f at $\mathbf{x} \in \mathbf{dom}(f)$ if*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{g}^\top(\mathbf{y} - \mathbf{x}) \; \forall \mathbf{y} \in \mathbf{dom}(f).$$

The set of all subgradients of f at $\mathbf{x}$ is called the subdifferential *and is denoted by ∂f.*

The subgradient is not always unique, hence the subdifferential set.

Lemma 1.2. *if f is convex and differentiable, then $\partial f(\mathbf{x}) = \{\nabla f(\mathbf{x}\}$.*

Which, to me, just comes from the fact that the derivative is unique. Similarly

Lemma 1.3. *If f is differentiable at $x \in \mathbf{dom}(f)$, then $\partial f(\mathbf{x}) \subseteq \{\nabla f(\mathbf{x})\}$.*

And there is a subgradient version of Lipschitz

Lemma 1.4. *Let f be convex with an open domain then*
i) $\|\mathbf{g}\| \leq B \; \forall \mathbf{x} \in \mathbf{dom}(f), \; \forall \mathbf{g} \in \partial f(\mathbf{x})$
ii) $|f\mathbf{y}) - f(\mathbf{x})| \leq B\|\mathbf{y} - \mathbf{x}\|$
are equivalent.

1.1 Topological properties

Of the subdifferential set that is.

Lemma 1.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then the subdifferential at any point $\mathbf{x} \in \mathbf{dom}(f)$ is a closed convex set.*

Now for something completely different

Definition 1.6. *Let S, T be two nonempty convex sets in $\mathbb{R}^n$ and let $H = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top\mathbf{x} = b\}, \; \mathbf{a} \neq 0$ is said to* separate *S and T if $S \cup T \not\subset H$ and*

$$S \subset H^- = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top\mathbf{x} \leq b\}$$
$$S \subset H^+ = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top\mathbf{x} \geq b\}$$

Similarly, the hyperplane is said to strictly separate *S and T if*

$$S \subset H^{--} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} < b\}$$
$$S \subset H^{++} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} > b\}$$

And then honestly some weird and seemingly useless shit about the subdifferential. Basically if f is convex then there are subdifferentials in the relative interior.

1.2 Subdifferential and directional derivative

Okay so now We are taking a derivative in a direction, namely

$$f'(\mathbf{x}, \mathbf{d}) = \lim_{\delta \to 0^+} \frac{f(x + \delta d) - f(x)}{\delta}$$

And when the function is differentiable We get $f'(\mathbf{x}, \mathbf{d}) = \nabla f(\mathbf{x})^\top \mathbf{d}$. I suppose this sort of makes sense - the derivative tells You the change in certain directions, the overall gradient vector then points in the direction of greatest change, and You can project that

Lemma 1.7. *When f is convex, the ratio*

$$\phi(\delta) = \frac{f(x + \delta d) - f(x)}{\delta}$$

is non-decreasing for $\delta > 0$.

Pretty simple, just means f is, well, non-decreasing (or rather it's gradient is).

I refuse to prove that the gradient is the direction of steepest ascent.

1.3 Calculus of Subgradient

So calculating the subdifferential is hard, actually, so the idea is to arrive at it constructively. Recall that We have that the subgradient exists almost everywhere for convex functions, so We can just compose them to get to more interesting cases.

We can take a linear combination of convex functions,

We can take an affine function and feed it into a convex function,

Taking the max of a set of convex functions,

And chain function apparently also works.

1.4 Subgradient method

Two quantities We'll care about are

$$R^2 = \max_{\mathbf{x},\mathbf{y}\in X} \|\mathbf{x}-\mathbf{y}\|^2$$

Where X is the subset of the domain of f that We care about. So this is just the squared diameter thanks.

and B, the Lipschitz constant. f will probably be locally Lipschitz.

1.5 Subgradient descent

Let $\mathbf{x}_1 \in X$ be the starting point, then

$$\mathbf{x}_{t+1} = \prod_X (\mathbf{x}_t - \gamma_t g(\mathbf{x}_t))$$

Theorem 1.8. *Assume f is convex, then Subgradient Descent satisfies*

$$\min_{1\geq t\geq T} f(\mathbf{x}_t) - f^* \leq \left(\sum_{t=1}^T \gamma_t\right)^{-1} \left(\frac{1}{2}\|\mathbf{x}_1-\mathbf{x}^*\|^2 + \frac{1}{2}\sum_{t=1}^T \gamma_t^2 \|g(\mathbf{x}_t)\|^2\right)$$

and

$$f(\hat{\mathbf{x}}_T) - f^* \leq \left(\sum_{t=1}^T \gamma_t\right)^{-1} \left(\frac{1}{2}\|\mathbf{x}_1-\mathbf{x}^*\|^2 + \frac{1}{2}\sum_{t=1}^T \gamma_t^2 \|g(\mathbf{x}_t)\|^2\right)$$

where

$$\hat{\mathbf{x}}_T = \left(\sum_{t=1}^T \gamma_t\right)^{-1} \left(\sum_{t=1}^T \gamma_\mathcal{T} \mathbf{x}_t\right)$$

Proof. Alright

$$\begin{aligned}
\|\mathbf{x}_{t+1}-\mathbf{x}^*\|^2 &= \|\prod_X (\mathbf{x}_t - \gamma_t g(\mathbf{x}_t)) - \prod_X (\mathbf{x}^*)\|^2 \\
&\leq \|\mathbf{x}_t - \gamma_t g(\mathbf{x}_t) - \mathbf{x}^*\|^2 \\
&\leq \|\mathbf{x}_t-\mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - 2(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*)
\end{aligned}$$

Where the first inequality is justified as follows: We are project two vectors into X and taking the distance between them. The projection operator can only decrease the distance between those vectors, since either they are not projected and the distance is unchanged or they are and they are brought closer (in the convex set X).

The second is just the cosine law for vectors.

Rearranging We get

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &\leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - 2(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) \\
2(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) &\leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \\
(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) &\leq \frac{1}{2}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)
\end{aligned}$$

Okay so next is that by convexity of f We have

$$(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) \geq \gamma_t (f(\mathbf{x}_t) - f^*)$$

How did We get here though. By convexity and subgradients We have

$$\begin{aligned}
f(\mathbf{y}) \geq f(\mathbf{x}) + g(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \implies \\
f(\mathbf{x}^*) \geq f(\mathbf{x}_t) + g(\mathbf{x}_t)^\top (\mathbf{x}^* - \mathbf{x}_t) \\
f(\mathbf{x}^*) - f(\mathbf{x}_t) \geq g(\mathbf{x}_t)^\top (\mathbf{x}^* - \mathbf{x}_t) \\
f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq g(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*) \\
\gamma_t (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \leq \gamma_t g(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*) \\
\gamma_t g(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*) \geq \gamma_t (f(\mathbf{x}_t) - f(\mathbf{x}^*)) \\
\gamma_t g(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*) \geq \gamma_t (f(\mathbf{x}_t) - f^*)
\end{aligned}$$

So that's annoying. But now We can write

$$\begin{aligned}
(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) &\leq \frac{1}{2}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big) \\
\gamma_t (f(\mathbf{x}_t) - f^*) &\leq \frac{1}{2}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big) \\
\sum_{t=1}^{T} \gamma_t (f(\mathbf{x}_t) - f^*) &\leq \sum_{t=1}^{T} \frac{1}{2}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big) \\
\sum_{t=1}^{T} \gamma_t (f(\mathbf{x}_t) - f^*) &\leq \frac{1}{2}\big(\|\mathbf{x}_1 - \mathbf{x}^*\|^2 - \|\mathbf{x}_{T+1} - \mathbf{x}^*\|^2 + \sum_{t=1}^{T} \|\gamma_t g(\mathbf{x}_t)\|^2\big) \\
\sum_{t=1}^{T} \gamma_t (f(\mathbf{x}_t) - f^*) &\leq \frac{1}{2}\big(\|\mathbf{x}_1 - \mathbf{x}^*\|^2 + \sum_{t=1}^{T} \|\gamma_t g(\mathbf{x}_t)\|^2\big)
\end{aligned}$$

We haven't actually done anything complicated yet. Just used the non-expansiveness of projection to get rid the projection operator, and replaced the gradient with a subgradient. In fact, I think this is identical to vanilla analysis except for those tricks.

Anyway, by definition, We have

$$\sum_{t=1}^{T}\gamma_t(f(\mathbf{x}_t)-f^*)\leq\sum_{t=1}^{T}\gamma_t\big(\min_{1\leq t\leq T}f(\mathbf{x}_t)-f^*\big)=\big(\min_{1\leq t\leq T}f(\mathbf{x}_t)-f^*\big)\sum_{t=1}^{T}\gamma_t$$

Which means We can say

$$\begin{aligned}\sum_{t=1}^{T}\gamma_t(f(\mathbf{x}_t)-f^*)&\leq\frac{1}{2}\big(\|\mathbf{x}_1-\mathbf{x}^*\|^2+\sum_{t=1}^{T}\|\gamma_t g(\mathbf{x}_t)\|^2\big)\\ \big(\min_{1\leq t\leq T}f(\mathbf{x}_t)-f^*\big)\sum_{t=1}^{T}\gamma_t&\leq\frac{1}{2}\big(\|\mathbf{x}_1-\mathbf{x}^*\|^2+\sum_{t=1}^{T}\|\gamma_t g(\mathbf{x}_t)\|^2\big)\\ \min_{1\leq t\leq T}f(\mathbf{x}_t)-f^*&\leq\Big(\sum_{t=1}^{T}\gamma_t\Big)^{-1}\frac{1}{2}\big(\|\mathbf{x}_1-\mathbf{x}^*\|^2+\sum_{t=1}^{T}\|\gamma_t g(\mathbf{x}_t)\|^2\big)\end{aligned}$$

Which proves the first direction. Wee.
For the other claim We use convexity. Basically We want to say

$$\sum_{t=1}^{T}\gamma_t(f(\mathbf{x}_t)-f^*)\geq\Big(\sum_{t=1}^{T}\gamma_t\Big)\cdot(f(\hat{\mathbf{x}}_T)-f^*)$$

And why is this true. Well, recall that

$$\hat{\mathbf{x}}_T=\Big(\sum_{t=1}^{T}\gamma_t\Big)^{-1}\Big(\sum_{t=1}^{T}\gamma_T\mathbf{x}_t\Big)$$

So $\hat{\mathbf{x}}_T$ is a convex sum as far as I can tell - the coefficients sum to 1. So We can arrive at our desired equation by starting (with convexity)

$$\begin{aligned}\Big(\sum_{t=1}^{T}\gamma_t\Big)^{-1}\sum_{t=1}^{T}\gamma_t f(\mathbf{x}_t)&\geq f\Big(\Big(\sum_{t=1}^{T}\gamma_t\Big)^{-1}\Big(\sum_{t=1}^{T}\gamma_t\mathbf{x}_t\Big)\Big)\\ \Big(\sum_{t=1}^{T}\gamma_t\Big)^{-1}\sum_{t=1}^{T}\gamma_t f(\mathbf{x}_t)&\geq f(\hat{\mathbf{x}}_T)\\ \sum_{t=1}^{T}\gamma_t f(\mathbf{x}_t)&\geq\Big(\sum_{t=1}^{T}\gamma_t\Big)f(\hat{\mathbf{x}}_T)\\ \sum_{t=1}^{T}\gamma_t f(\mathbf{x}_t-f^*)&\geq\Big(\sum_{t=1}^{T}\gamma_t\Big)(f(\hat{\mathbf{x}}_T)-f^*)\end{aligned}$$

Where in that last step We just subtracted $\sum_{t=1}^{T} \gamma_t f^*$ to get the desired form.

To complete the proof We go back to our original formulation:

$$\sum_{t=1}^{T} \gamma_t (f(\mathbf{x}_t) - f^*) \leq \frac{1}{2}\big(\|\mathbf{x}_1 - \mathbf{x}^*\|^2 + \sum_{t=1}^{T} \|\gamma_t g(\mathbf{x}_t)\|^2\big)$$

$$\Big(\sum_{t=1}^{T} \gamma_t\Big)(f(\hat{\mathbf{x}}_T) - f^*) \leq \sum_{t=1}^{T} \gamma_t (f(\mathbf{x}_t) - f^*) \leq \frac{1}{2}\big(\|\mathbf{x}_1 - \mathbf{x}^*\|^2 + \sum_{t=1}^{T} \|\gamma_t g(\mathbf{x}_t)\|^2\big)$$

$$\Big(\sum_{t=1}^{T} \gamma_t\Big)(f(\hat{\mathbf{x}}_T) - f^*) \leq \frac{1}{2}\big(\|\mathbf{x}_1 - \mathbf{x}^*\|^2 + \sum_{t=1}^{T} \|\gamma_t g(\mathbf{x}_t)\|^2\big)$$

$$f(\hat{\mathbf{x}}_T) - f^* \leq \Big(\sum_{t=1}^{T} \gamma_t\Big)^{-1} \frac{1}{2}\big(\|\mathbf{x}_1 - \mathbf{x}^*\|^2 + \sum_{t=1}^{T} \|\gamma_t g(\mathbf{x}_t)\|^2\big)$$

□

Observation 1.9. *By the way, You can change the previous theorem to read*

$$\min_{1\leq t\leq T} f(\mathbf{x}_t) - f^* \leq \frac{\frac{1}{2}\big(\|\mathbf{x}_1 - \mathbf{x}^*\|^2 + \sum_{t=1}^{T} \|\gamma_t g(\mathbf{x}_t)\|^2\big)}{\sum_{t=1}^{T} \gamma_t}$$

$$\min_{1\leq t\leq T} f(\mathbf{x}_t) - f^* \leq \frac{\frac{1}{2}\big(R^2 + \sum_{t=1}^{T} \gamma_t^2 B^2\big)}{\sum_{t=1}^{T} \gamma_t}$$

By letting R be distance to solution and B be some sort of lipschitz upper bound.

Now for some **convergence with various stepsizes**:

1. Constant stepsize $\gamma_t = t$:

We let ϵ_T be the error after T steps, and We see what happens to it:

$$\epsilon_T \leq \frac{\frac{1}{2}\big(R^2 + \sum_{t=1}^{T} \gamma_t^2 B^2\big)}{\sum_{t=1}^{T} \gamma_t}$$

$$\epsilon_T \leq \frac{\frac{1}{2}\big(R^2 + \sum_{t=1}^{T} \gamma^2 B^2\big)}{\sum_{t=1}^{T} \gamma}$$

$$\epsilon_T \leq \frac{R^2 + T\gamma^2 B^2}{2T\gamma}$$

$$\epsilon_T \leq \frac{R^2}{2T\gamma} + \frac{T\gamma^2 B^2}{2T\gamma}$$

$$\epsilon_T \leq \frac{R^2}{2T\gamma} + \frac{\gamma B^2}{2}$$

$$\epsilon_T \leq \frac{\gamma B^2}{2} \quad \lim_{T\to\infty}$$

So the error does not diminish to zero even with an infinite number of steps, which makes sense. You can of course also take the derivative of the right side and try to find the optimal step size, yielding $\gamma^* = R/(B\sqrt{T})$.

2. Non-summable but diminishing step size

So We have that

$$\sum_{t=1}^{T} \gamma_t = \infty, \quad \lim_{T\to\infty} \gamma_t = 0$$

$$\epsilon_T \leq \frac{\frac{1}{2}\left(R^2 + \sum_{t=1}^{T} \gamma_t^2 B^2\right)}{\sum_{t=1}^{T} \gamma_t}$$
$$\epsilon_T \leq \frac{R^2}{2\sum_{t=1}^{T} \gamma_t} + \frac{2\sum_{t=1}^{T} \gamma_t^2 B^2}{\sum_{t=1}^{T} \gamma_t}$$

Well, that first term goes to zero clearly. Similarly on the right hand side the numerator goes to zero faster then the denumerator.

3. Non-summable but square summable:

So sum of the first order terms is infinite, sum of squared is bounded. That's very clearly going to go to zero as per previous example.

4. Polyak is, well, I don't care, but goes to zero.

1.6 Convergence for strongly convex functions

Theorem 1.10. *Assume f is μ-strongly convex, then subgradient descent with step size*

$$\gamma_t = \frac{1}{\mu t}$$

satisfies

$$\min_{1\leq t\leq T} f(\mathbf{x}_t) - f^* \leq \frac{B^2 \log(T) + 1)}{2\mu T}$$

and

$$f(\hat{\mathbf{x}}_T) - f^* \leq \frac{B^2 \log(T) + 1)}{2\mu T},$$

where $\hat{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$.

Proof. First We have strong convexity:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

then let $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^*$:

$$
\begin{aligned}
f(\mathbf{x}^*) &\geq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}^* - \mathbf{x}_t) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2 \\
f(\mathbf{x}^*) &\geq f(\mathbf{x}_t) + \mathbf{g}_t^\top (\mathbf{x}^* - \mathbf{x}_t) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2 \\
\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*) &\geq f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2 \\
\gamma_t \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*) &\geq \gamma_t \left(f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2 \right)
\end{aligned}
$$

Recall We had

$$
\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &= \| \prod_X (\mathbf{x}_t - \gamma_t g(\mathbf{x}_t)) - \prod_X (\mathbf{x}^*) \|^2 \\
&\leq \|\mathbf{x}_t - \gamma_t g(\mathbf{x}_t) - \mathbf{x}^*\|^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - 2(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*)
\end{aligned}
$$

and by rearranging that We had

$$
\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 &\leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - 2(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) \\
2(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) &\leq \|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2 \\
(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) &\leq \frac{1}{2}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)
\end{aligned}
$$

So now We just stick our new strongly convex result in there

$$
\begin{aligned}
(\gamma_t g(\mathbf{x}_t))^\top (\mathbf{x}_t - \mathbf{x}^*) &\leq \frac{1}{2}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big) \\
\gamma_t\big(f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2\big) &\leq \\
\frac{1}{2}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big) & \\
f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2 &\leq \frac{1}{2\gamma_t}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big) \\
f(\mathbf{x}_t) - f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2 &\leq \frac{\mu t}{2}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)
\end{aligned}
$$

Fuck that. We juts subtract $\frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2$ from both stupid sides to get

$$
\begin{aligned}
&-\frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_t\|^2 + \frac{\mu t}{2}\big(\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)\\
&\frac{\mu(t-1)}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \frac{\mu t}{2}\big(\|\gamma_t g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)\\
&\frac{\mu(t-1)}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \frac{\mu t}{2}\big(\gamma_t^2\|g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)\\
&\frac{\mu(t-1)}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \frac{\mu t}{2}\big(\frac{4}{\mu^2 t^2}\|g(\mathbf{x}_t)\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)\\
&\frac{\mu(t-1)}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \frac{2}{\mu t}\|g(\mathbf{x}_t)\|^2 - \frac{\mu t}{2}\big(\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)
\end{aligned}
$$

Ugh.

We therefore have

$$
f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{\mu(t-1)}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 + \frac{2}{\mu t}\|g(\mathbf{x}_t)\|^2 - \frac{\mu t}{2}\big(\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)
$$

Then it's a sum but I am confused about the coefficients. The gradient term We can take out since there is no cancellation there.

So We have

$$
\begin{aligned}
\frac{\mu(t-1)}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \frac{\mu t}{2}\big(\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)\\
\frac{\mu t}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \frac{\mu t}{2}\big(\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\big)\\
\frac{\mu}{2}\bigg(t\|\mathbf{x}_t - \mathbf{x}^*\|^2 - \|\mathbf{x}_t - \mathbf{x}^*\|^2 - t\big(\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\bigg)\\
\frac{\mu}{2}\bigg((t-1)\|\mathbf{x}_t - \mathbf{x}^*\|^2 - t\big(\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2\bigg)
\end{aligned}
$$

And so now the question is does this telescope. It does. The first term remains, then You get t of the second term. At $t+1$ You get $t+1-1$ and it all works out. The very first term also disappears since at $t = 1$ You're left with $t - 1 = 0$ factor.

You also get a $-T\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2$, but since this is a strictly negative quantity We can drop it. This yields

$$
\begin{aligned}
\sum_{t=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \sum_{t=1}^{T} \frac{2}{\mu t}\|g(\mathbf{x}_t)\|^2\\
\sum_{t=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\leq \frac{2B^2}{\mu t}\sum_{t=1}^{T}\frac{1}{t}
\end{aligned}
$$

And then apparently it's just a known fact that

$$\sum_{t=1}^{T} \frac{1}{t} \leq \log(T) + 1$$

which gives

$$\sum_{t=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{2B^2}{\mu t}(\log(T) + 1)$$

And the rest is as before, minimum is less than any of the terms on the right, or take a linear combination of the inputs.

□