ECE826 Lecture 7:

How fast is Gradient Descent?

Contents

- Convergence Rates
- Gradient Descent
- GD on smooth U lipschitz U str. convex objectives
- Complexity of GD

Minimizing the Empirical Risk

- The empirical cost function that we have access to

$$\min_{h \in \mathscr{H}} \left(R_S[h] = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i); y_i) \right)$$

- Question: Can we approximate the solution to this minimization? If so how fast?

- The answer must depend on:
  1) n, the sample size
  2) $\mathscr{H}$, the hypothesis class and loss function
  3) $\mathscr{D}$, the data distribution
  4) the optimization algorithm that outputs our classifier

Computational Aspects of the ERM

Last time

- ERM is hard
- Learning & memorizing is hard for fixed architecture
- Memorizing is easy, assuming arbitrary architecture
- Convexity can help, but by how much?

Stop 1: Convexity

First stop: Convexity

- "A function that looks like a bowl"

Def.:

A function $f(w)$ is convex on $\mathscr{W}$ if

$$f(a \cdot w + (1-a) \cdot w') \leq af(w) + (1-a)f(w')$$

First stop: Convexity

- "A function that looks like a bowl"

Def.:

A function $f(w)$ is convex on $\mathscr{W}$ if

$$f(a \cdot w + (1-a) \cdot w') \leq af(w) + (1-a)f(w')$$

- Convexity makes our lives much easier

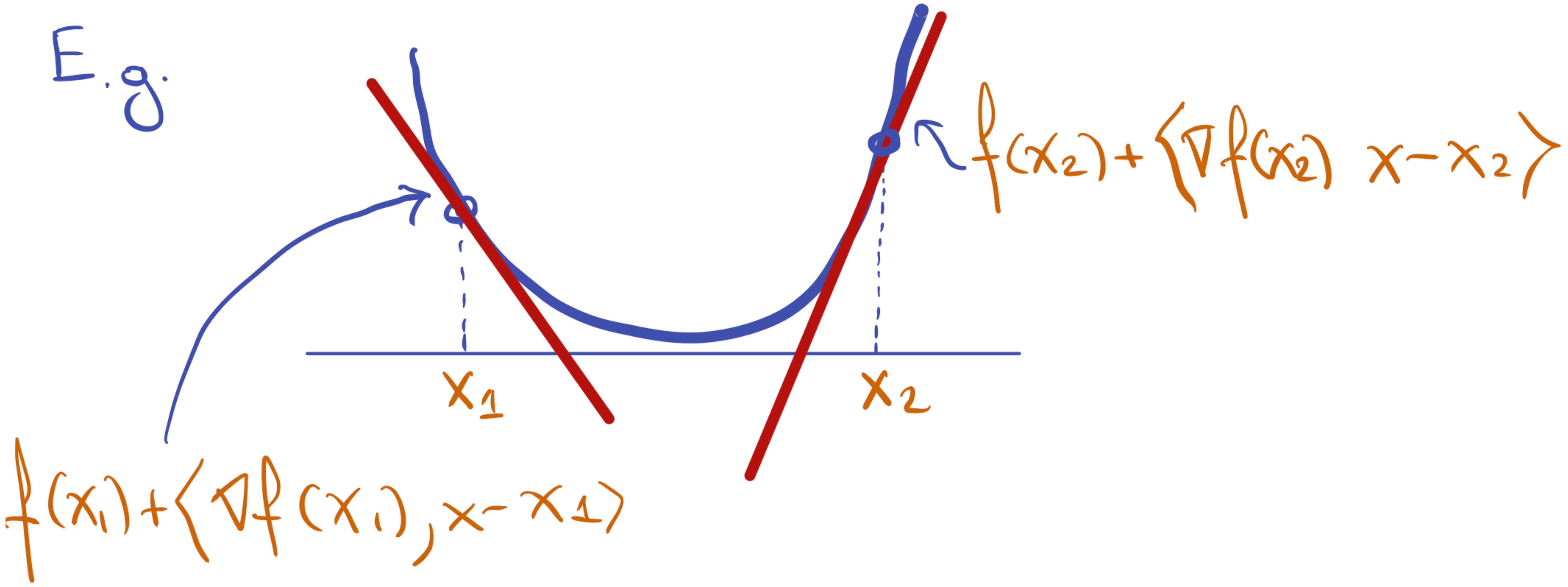
$$\langle \nabla f(w'), w' - w^* \rangle \geq f(w') - f(w^*)$$

gradient is always positively correlated with the right direction towards OPT

Let's get a bit more mileage from this

First stop: Convexity

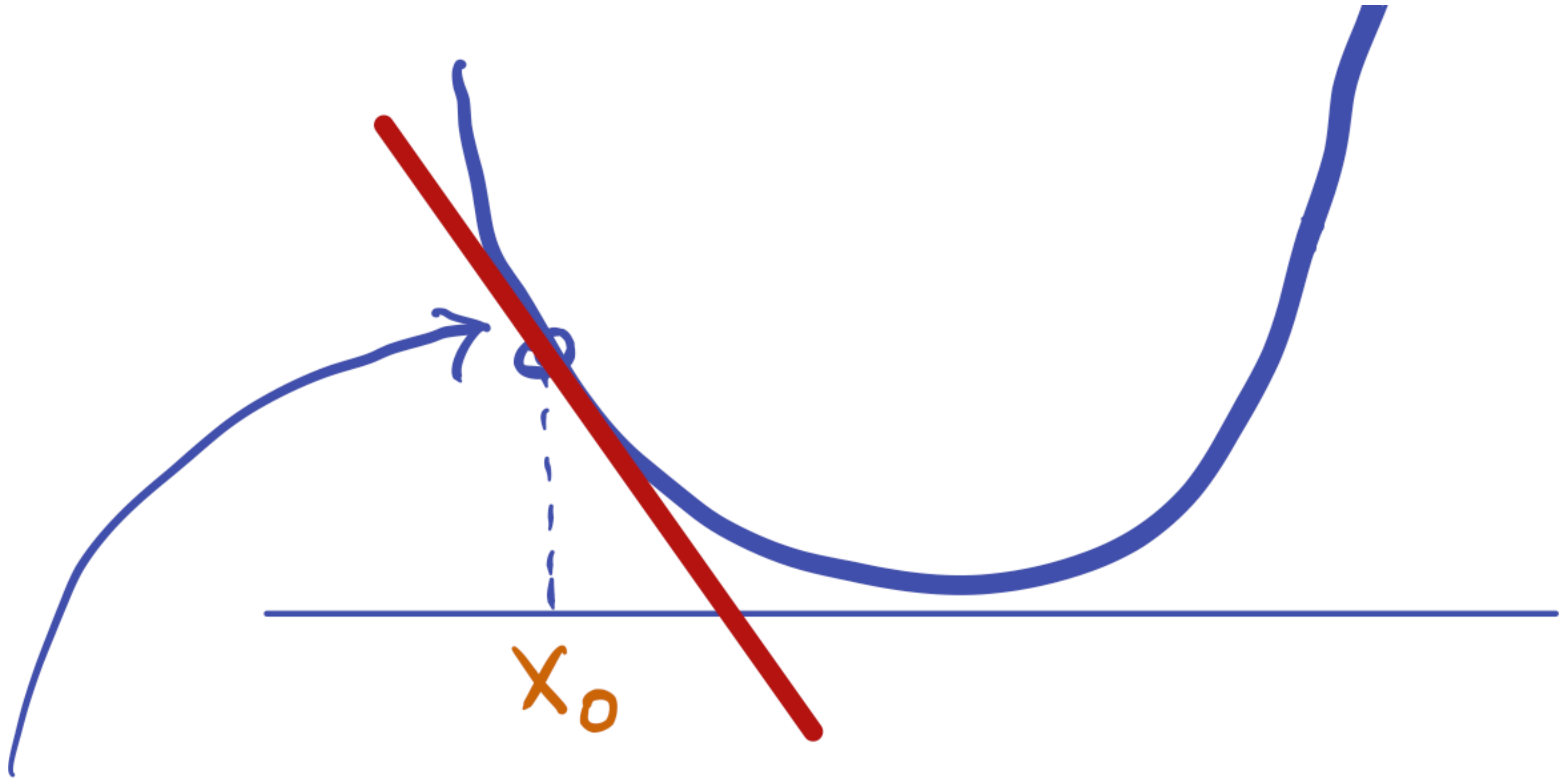
- The first order Taylor expansion of a convex function is a "global under-estimate"
- $\forall w, w_0 \in \mathbb{R}^d,\ \ f(w) \geq f(w_0) + \langle \nabla f(w_0), w - w_0 \rangle$



- Observe: 1-st order Taylor always has a linear form, e.g. , $f(w) \approx \langle w, a \rangle + b$

Q: what happens for $w_0\ \ s.t.\ \ \nabla f(w_0) = 0$?

Local optimization



- Say we initialize at w_0, then we could try to follow the "line" $f(w_0) + \langle f(w_0), w - w_0 \rangle$!
- Let's make our algorithm to progress by additive steps, i.e.,

$$w_{k+1} = \arg\min_{w \in \mathscr{W}} \left\{ f(w_k) + \langle \nabla f(w_k, w - w_k \rangle + \frac{1}{2\gamma} \|w - w_k\|^2 \right\}$$

The GD step is the solution to the above:

$$w_{k+1} = w_k - \gamma \nabla f(w_k)$$

Does GD converge, and if so how fast?

Convergence rates?

- Convergence rates = a promise of worst case performance not being bad

$$f(w_T) - f(w^*) \leq \rho(T, f, w_0)$$

Convergence rates?

- Convergence rates = a promise of worst case performance not being bad

$$f(w_T) - f(w^*) \leq \rho(T, f, w_0)$$

You would typically like $\rho(T, f, w_0) \sim \dfrac{1}{\text{poly}(T)}$

Convergence rates?

- Convergence rates = a promise of worst case performance not being bad

$$f(w_T) - f(w^*) \leq \rho(T, f, w_0)$$

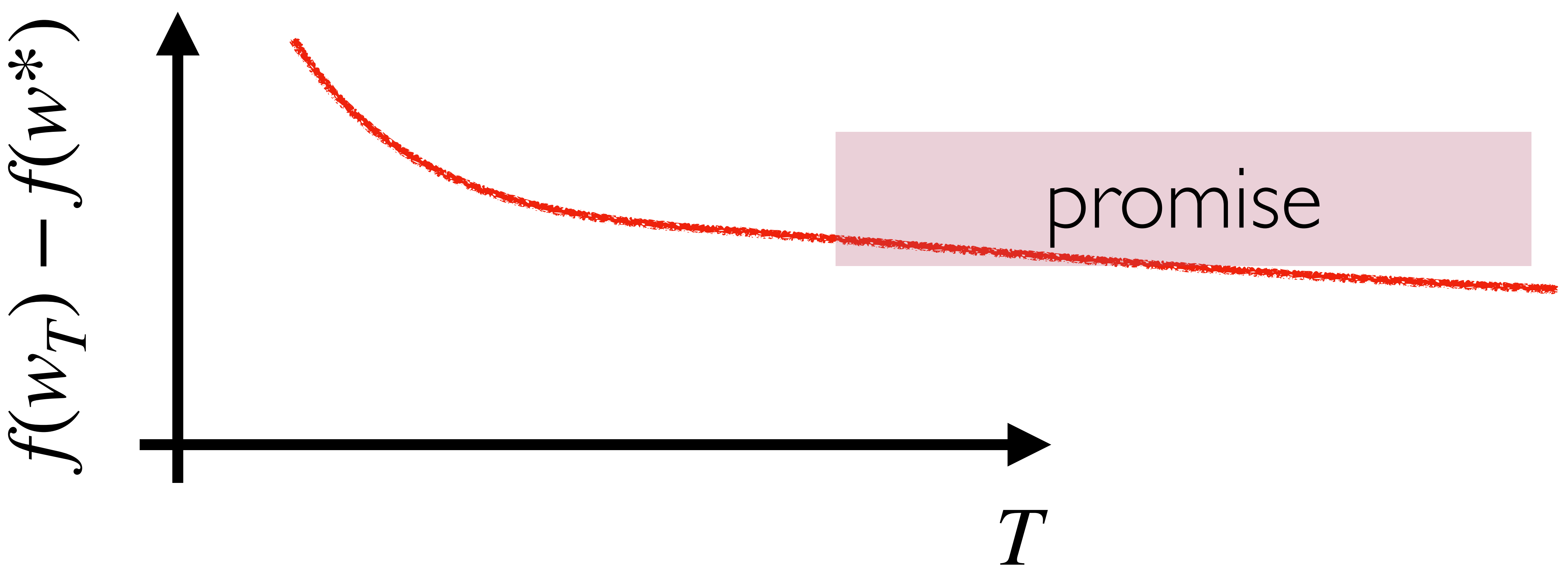
You would typically like $\rho(T, f, w_0) \sim \dfrac{1}{\text{poly}(T)}$

Convergence rates?

- Convergence rates = a promise of worst case performance not being bad

$$f(w_T) - f(w^*) \leq \rho(T, f, w_0)$$

You would typically like $\rho(T, f, w_0) \sim \dfrac{1}{\text{poly}(T)}$

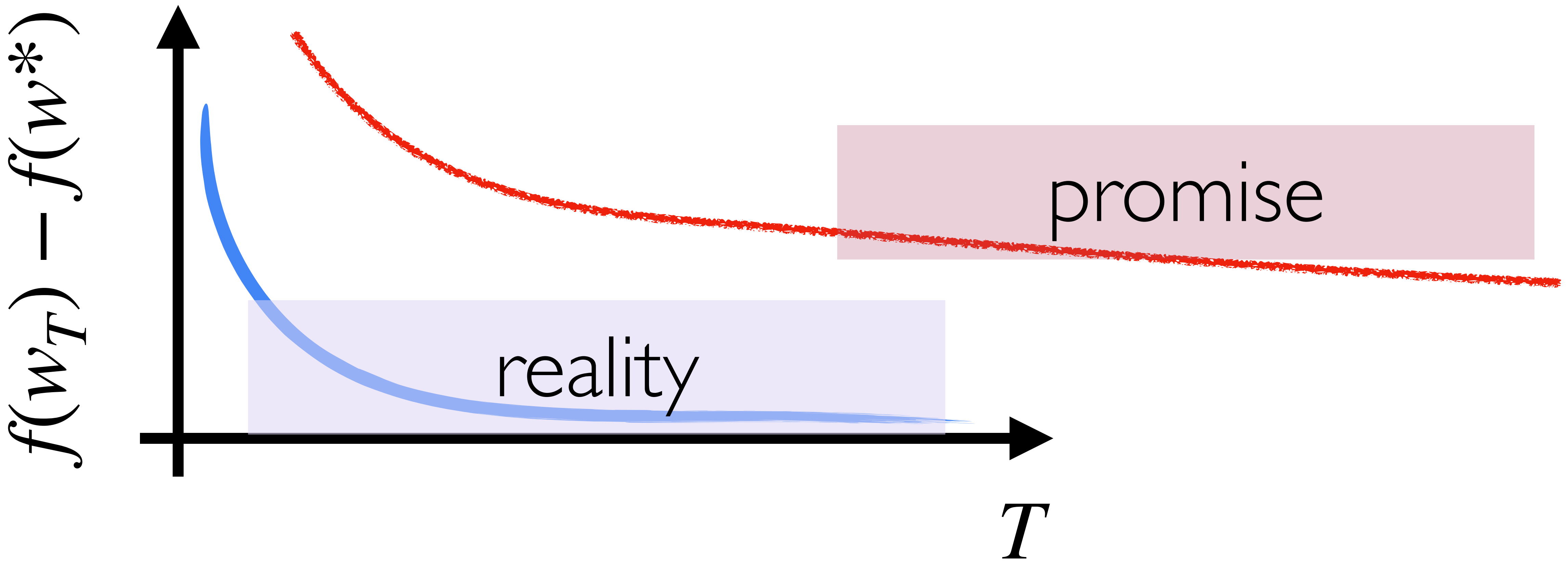


Convergence rates?

- Convergence rates = a promise of worst case performance not being bad

$$f(w_T) - f(w^*) \leq \rho(T, f, w_0)$$

You would typically like $\rho(T, f, w_0) \sim \dfrac{1}{\text{poly}(T)}$

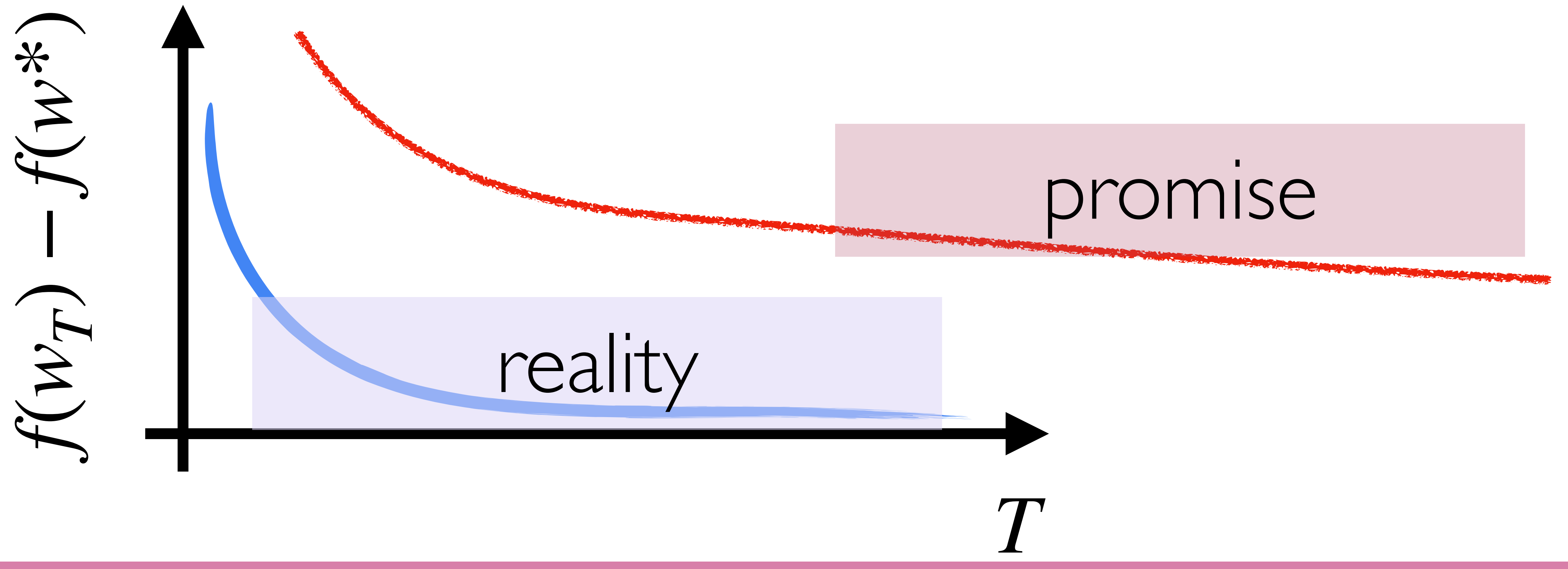


Convergence rates?

- Convergence rates = a promise of worst case performance not being bad

$$f(w_T) - f(w^*) \le \rho(T, f, w_0)$$

You would typically like $\rho(T, f, w_0) \sim \dfrac{1}{\text{poly}(T)}$



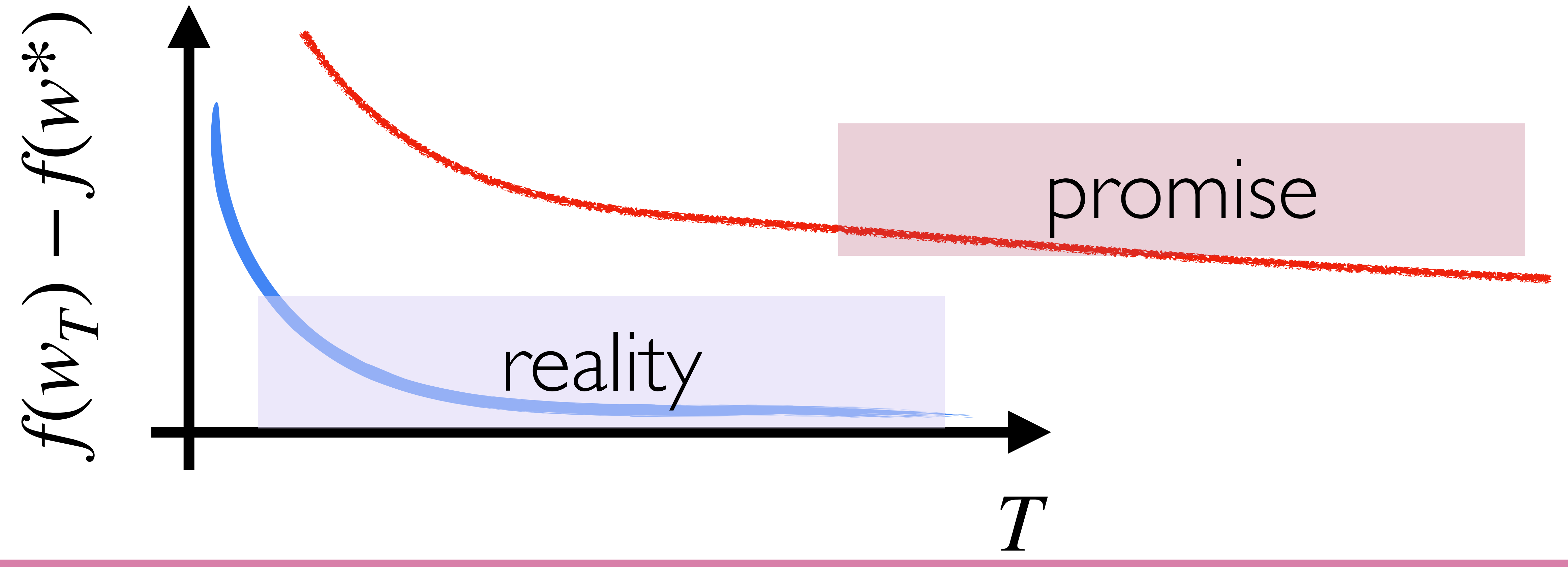
- Warning 1: Worst case bounds may be too pessimistic and not too close to reality.
- Warning 2: If A has faster convergence rate than B, doesn't mean A is faster than B in practice

Convergence rates?

- Convergence rates = a promise of worst case performance not being bad

$$f(w_T) - f(w^*) \leq \rho(T, f, w_0)$$

You would typically like $\rho(T, f, w_0) \sim \dfrac{1}{\text{poly}(T)}$



- Warning 1: Worst case bounds may be too pessimistic and not too close to reality.
- Warning 2: If A has faster convergence rate than B, doesn't mean A is faster than B in practice

Convergence rates can be informative in understanding what structures allow for faster algorithms and can be good guides towards algorithm design.

GD on Lipschitz + CVX functions

Lipschitzness

- Lipschitz: "A function can't change too fast"

Def.:

A function $f(w)$ is L-Lipschitz on $\mathscr{W}$ if

$$|f(w) - f(w')| \leq L \cdot \|w - w'\|, \forall w, w' \in \mathscr{W}$$

Lipschitzness is implied by the property $\forall w \in \mathscr{W},\ \|\nabla f(w)\| \leq L$ which we will assume

How to show convergence?

- Let's start with the "under-estimator" property of convexity

$$f(w_k) - f(w^\star) \leq \langle \nabla f(w_k), w_k - w^\star \rangle = \left\langle \frac{w_k - w_{k+1}}{\gamma}, w_k - w^\star \right\rangle .$$

How to show convergence?

- Let's start with the "under-estimator" property of convexity

$$f(w_k) - f(w^\star) \leq \langle \nabla f(w_k), w_k - w^\star \rangle = \left\langle \frac{w_k - w_{k+1}}{\gamma}, w_k - w^\star \right\rangle.$$

- We would like instead of the inner products of iterate differences to work with norms (this will be related to the Lip. continuity)

How to show convergence?

- Let's start with the "under-estimator" property of convexity

$$f(w_k) - f(w^\star) \leq \langle \nabla f(w_k), w_k - w^\star \rangle = \left\langle \frac{w_k - w_{k+1}}{\gamma}, w_k - w^\star \right\rangle.$$

- We would like instead of the inner products of iterate differences to work with norms (this will be related to the Lip. continuity)

- Then, remember that $a^T b = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a - b\|^2\right)$

How to show convergence?

- Let's start with the "under-estimator" property of convexity

$$f(w_k) - f(w^\star) \leq \langle \nabla f(w_k), w_k - w^\star \rangle = \left\langle \frac{w_k - w_{k+1}}{\gamma}, w_k - w^\star \right\rangle .$$

- We would like instead of the inner products of iterate differences to work with norms (this will be related to the Lip. continuity)

- Then, remember that $a^T b = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a - b\|^2 \right)$

- Which implies

$$f(w_k) - f(x^\star) \leq \frac{1}{2\gamma}\left\{ \|w_k - w^\star\|^2 + \|w_k - w_{k+1}\|^2 - \|w_{k+1} - w^\star\|^2 \right\}$$

How to show convergence?

- Let's start with the "under-estimator" property of convexity

$$f(w_k) - f(w^\star) \leq \langle \nabla f(w_k), w_k - w^\star \rangle = \left\langle \frac{w_k - w_{k+1}}{\gamma}, w_k - w^\star \right\rangle.$$

- We would like instead of the inner products of iterate differences to work with norms (this will be related to the Lip. continuity)

- Then, remember that $a^T b = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a - b\|^2\right)$

- Which implies

$$\begin{aligned} f(w_k) - f(x^\star) &\leq \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 + \|w_k - w_{k+1}\|^2 - \|w_{k+1} - w^\star\|^2\right\} \\ &= \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 + \|\gamma \nabla f(w_k)\|^2 - \|w_{k+1} - w^\star\|^2\right\} \end{aligned}$$

How to show convergence?

- Let's start with the "under-estimator" property of convexity

$$f(w_k) - f(w^\star) \le \langle \nabla f(w_k), w_k - w^\star \rangle = \left\langle \frac{w_k - w_{k+1}}{\gamma}, w_k - w^\star \right\rangle .$$

- We would like instead of the inner products of iterate differences to work with norms (this will be related to the Lip. continuity)

- Then, remember that $a^T b = \frac{1}{2}\left(\|a\|^2 + \|b\|^2 - \|a-b\|^2\right)$

- Which implies

$$\begin{aligned} f(w_k) - f(x^\star) &\le \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 + \|w_k - w_{k+1}\|^2 - \|w_{k+1} - w^\star\|^2\right\} \\ &= \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 + \|\gamma \nabla f(w_k)\|^2 - \|w_{k+1} - w^\star\|^2\right\} \\ &\le \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 - \|w_{k+1} - w^\star\|^2\right\} + \frac{\gamma L^2}{2} \end{aligned}$$

Convergence rates?

- Let's calculate the sub optimality gap for all steps:

$$f(w_k) - f(w^\star) \le \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 - \|w_{k+1} - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

Convergence rates?

- Let's calculate the sub optimality gap for all steps:

$$f(w_k) - f(w^\star) \leq \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 - \|w_{k+1} - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

$$f(w_{k-1}) - f(w^\star) \leq \frac{1}{2\gamma}\left\{\|w_{k-1} - w^\star\|^2 - \|w_k - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

Convergence rates?

- Let's calculate the sub optimality gap for all steps:

$$
\begin{aligned}
f(w_k) - f(w^\star) &\leq \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 - \|w_{k+1} - w^\star\|^2\right\} + \frac{\gamma L^2}{2} \\
f(w_{k-1}) - f(w^\star) &\leq \frac{1}{2\gamma}\left\{\|w_{k-1} - w^\star\|^2 - \|w_k - w^\star\|^2\right\} + \frac{\gamma L^2}{2} \\
&\vdots \\
f(w_0) - f(w^\star) &\leq \frac{1}{2\gamma}\left\{\|w_0 - w^\star\|^2 - \|w_1 - w^\star\|^2\right\} + \frac{\gamma L^2}{2}
\end{aligned}
$$

Convergence rates?

- Let's calculate the sub optimality gap for all steps:

$$f(w_k) - f(w^\star) \leq \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 - \|w_{k+1} - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

$$f(w_{k-1}) - f(w^\star) \leq \frac{1}{2\gamma}\left\{\|w_{k-1} - w^\star\|^2 - \|w_k - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

$$\vdots$$

$$f(w_0) - f(w^\star) \leq \frac{1}{2\gamma}\left\{\|w_0 - w^\star\|^2 - \|w_1 - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

Convergence rates?

- Let's calculate the sub optimality gap for all steps:

$$f(w_k) - f(w^\star) \leq \frac{1}{2\gamma}\{\|w_k - w^\star\|^2 - \|w_{k+1} - w^\star\|^2\} + \frac{\gamma L^2}{2}$$

$$f(w_{k-1}) - f(w^\star) \leq \frac{1}{2\gamma}\{\|w_{k-1} - w^\star\|^2 - \|w_k - w^\star\|^2\} + \frac{\gamma L^2}{2}$$

$$\vdots$$

$$f(w_0) - f(w^\star) \leq \frac{1}{2\gamma}\{\|w_0 - w^\star\|^2 - \|w_1 - w^\star\|^2\} + \frac{\gamma L^2}{2}$$

If you add all these inequalities the highlighted terms go away!

Convergence rates?

- Let's calculate the sub optimality gap for all steps:

$$f(w_k) - f(w^\star) \leq \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 - \|w_{k+1} - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

$$f(w_{k-1}) - f(w^\star) \leq \frac{1}{2\gamma}\left\{\|w_{k-1} - w^\star\|^2 - \|w_k - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

$$\vdots$$

$$f(w_0) - f(w^\star) \leq \frac{1}{2\gamma}\left\{\|w_0 - w^\star\|^2 - \|w_1 - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

- Now sum them all up!

Convergence rates?

- Let's calculate the sub optimality gap for all steps:

$$f(w_k) - f(w^\star) \le \frac{1}{2\gamma}\left\{\|w_k - w^\star\|^2 - \|w_{k+1} - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

$$f(w_{k-1}) - f(w^\star) \le \frac{1}{2\gamma}\left\{\|w_{k-1} - w^\star\|^2 - \|w_k - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

$$\vdots$$

$$f(w_0) - f(w^\star) \le \frac{1}{2\gamma}\left\{\|w_0 - w^\star\|^2 - \|w_1 - w^\star\|^2\right\} + \frac{\gamma L^2}{2}$$

- Now sum them all up!

$$\sum_{t=1}^{T}\left(f(w_t) - f(w^\star)\right) \le \frac{-\|w_{T+1} - w^\star\|^2 + \|w_0 - w^\star\|^2}{2\gamma} + T\frac{\gamma L^2}{2}$$

Convergence rates?

- One more step..

$$\sum_{t=1}^{T} \left(f(w_t) - f(w^\star)\right) \leq \frac{-\|w_{T+1} - w^\star\|^2 + \|w_0 - w^\star\|^2}{2\gamma} + T\frac{\gamma L^2}{2}$$

Convergence rates?

- One more step..

$$\sum_{t=1}^{T} \left(f(w_t) - f(w^\star)\right) \leq \frac{-\|w_{T+1} - w^\star\|^2 + \|w_0 - w^\star\|^2}{2\gamma} + T\frac{\gamma L^2}{2}$$

$$\implies \frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^\star) \leq \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

Convergence rates?

- One more step..

$$\sum_{t=1}^{T} \left(f(w_t) - f(w^\star)\right) \leq \frac{-\|w_{T+1} - w^\star\|^2 + \|w_0 - w^\star\|^2}{2\gamma} + T\frac{\gamma L^2}{2}$$

$$\implies \frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^\star) \leq \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

- Due to convexity we have

Convergence rates?

- One more step..

$$\sum_{t=1}^{T} \left(f(w_t) - f(w^\star)\right) \leq \frac{-\|w_{T+1} - w^\star\|^2 + \|w_0 - w^\star\|^2}{2\gamma} + T\frac{\gamma L^2}{2}$$

$$\implies \frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^\star) \leq \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

- Due to convexity we have

$$f\left(\frac{1}{T}\sum_{i=1}^{T} w_k\right) - f(w^\star) \leq \frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^\star) \leq \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

Convergence rates?

- One more step..

$$\sum_{t=1}^{T}\left(f(w_t) - f(w^\star)\right) \le \frac{-\|w_{T+1} - w^\star\|^2 + \|w_0 - w^\star\|^2}{2\gamma} + T\frac{\gamma L^2}{2}$$

$$\implies \frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^\star) \le \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

- Due to convexity we have

$$f\left(\frac{1}{T}\sum_{i=1}^{T} w_k\right) - f(w^\star) \le \frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w^\star) \le \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

the sum of
iterates…

Convergence rates?

- One more step..

$$\sum_{t=1}^{T}\left(f(w_t)-f(w^\star)\right) \leq \frac{-\|w_{T+1}-w^\star\|^2+\|w_0-w^\star\|^2}{2\gamma} + T\frac{\gamma L^2}{2}$$

$$\implies \frac{1}{T}\sum_{t=1}^{T} f(w_t)-f(w^\star) \leq \frac{\|w_0-w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

- Due to convexity we have

$$f\left(\frac{1}{T}\sum_{i=1}^{T} w_k\right) - f(w^\star) \leq \frac{1}{T}\sum_{t=1}^{T} f(w_t)-f(w^\star) \leq \frac{\|w_0-w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

the sum of iterates…

almost what i want, but that additive term is bothering me

Convergence rates?

- One more step..

$$\sum_{t=1}^{T}\left(f(w_t)-f(w^\star)\right) \le \frac{-\|w_{T+1}-w^\star\|^2+\|w_0-w^\star\|^2}{2\gamma}+T\frac{\gamma L^2}{2}$$

$$\implies \frac{1}{T}\sum_{t=1}^{T} f(w_t)-f(w^\star) \le \frac{\|w_0-w^\star\|^2}{2\gamma T}+\frac{\gamma L^2}{2}$$

- Due to convexity we have

$$f\left(\frac{1}{T}\sum_{i=1}^{T} w_k\right)-f(w^\star) \le \frac{1}{T}\sum_{t=1}^{T} f(w_t)-f(w^\star) \le \frac{\|w_0-w^\star\|^2}{2\gamma T}+\frac{\gamma L^2}{2}$$

the sum of iterates…

almost what i want, but that [illegible] term is bothering me

now we get to choose… the step-size!

Convergence rates?

- We would like both of the terms on the right to be of the same order

$$f\left(\frac{1}{T}\sum_{i=1}^{T} w_k\right) - f(w^\star) \le \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

Convergence rates?

- We would like both of the terms on the right to be of the same order

$$f\left(\frac{1}{T}\sum_{i=1}^{T} w_k\right) - f(w^\star) \leq \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

Convergence rates?

- We would like both of the terms on the right to be of the same order

$$f\Big(\frac{1}{T}\sum_{i=1}^{T} w_k\Big) - f(w^\star) \leq \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

- Let's try to minimize them as a function of the step-size

Convergence rates?

- We would like both of the terms on the right to be of the same order

$$f\Big(\frac{1}{T}\sum_{i=1}^{T} w_k\Big) - f(w^\star) \leq \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

- Let's try to minimize them as a function of the step-size

$$\min_{\gamma} \frac{R^2}{2\gamma T} + \frac{\gamma L^2}{2} \implies \gamma = \frac{R}{L\sqrt{T}}\,.$$

Convergence rates?

- We would like both of the terms on the right to be of the same order

$$f\Big(\frac{1}{T}\sum_{i=1}^{T} w_k\Big) - f(w^\star) \le \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

- Let's try to minimize them as a function of the step-size

$$\min_\gamma \frac{R^2}{2\gamma T} + \frac{\gamma L^2}{2} \implies \gamma = \frac{R}{L\sqrt{T}}.$$

- which will leads to

$$f\Big(\frac{1}{T}\sum_{i=1}^{T} x_k\Big) - f(x^\star) \le \frac{RL}{\sqrt{T}}.$$

Convergence rates?

- We would like both of the terms on the right to be of the same order

$$f\left(\frac{1}{T}\sum_{i=1}^{T} w_k\right) - f(w^\star) \le \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

- Let's try to minimize them as a function of the step-size

$$\min_{\gamma} \frac{R^2}{2\gamma T} + \frac{\gamma L^2}{2} \implies \gamma = \frac{R}{L\sqrt{T}}.$$

- which will leads to

$$f\left(\frac{1}{T}\sum_{i=1}^{T} x_k\right) - f(x^\star) \le \frac{RL}{\sqrt{T}}.$$

We're done! GD on Lip+CVX functions converges at a rate of $\sim \frac{1}{\sqrt{T}}$

Convergence rates?

- We would like both of the terms on the right to be of the same order

$$f\Big(\frac{1}{T}\sum_{i=1}^{T} w_k\Big) - f(w^\star) \le \frac{\|w_0 - w^\star\|^2}{2\gamma T} + \frac{\gamma L^2}{2}$$

- Let's try to minimize them as a function of the step-size

$$\min_\gamma \frac{R^2}{2\gamma T} + \frac{\gamma L^2}{2} \implies \gamma = \frac{R}{L\sqrt{T}}.$$

- which will leads to

$$f\Big(\frac{1}{T}\sum_{i=1}^{T} x_k\Big) - f(x^\star) \le \frac{RL}{\sqrt{T}}.$$

Q: how many steps for ε approx?

We're done! GD on Lip+CVX functions converges at a rate of $\sim \frac{1}{\sqrt{T}}$

GD on smooth + str. CVX functions

Str. Convexity & smoothness

Def.:

A function $f(w)$ is λ-strongly convex on $\mathscr{W}$ if

$$f(w) \geq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{\lambda}{2}\|w - w'\|^2$$

- The best kind of convexity, aka $f(w) \geq f(w^*) + \frac{\lambda}{2}\|w - w^*\|^2$

Def.:

A function $f(w)$ is β-Lipschitz on $\mathscr{W}$ if

$$\|\nabla f(w) - \nabla f(w')\| \leq \beta \cdot \|w - w'\|, \forall w, w' \in \mathscr{W}$$

- A quadratic upper bound $f(w) \leq f(w^*) + \frac{\beta}{2}\|w - w^*\|^2$

Str. Convexity & smoothness

Def.:

A function $f(w)$ is λ-strongly convex on $\mathscr{W}$ if

$$f(w) \geq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{\lambda}{2}\|w - w'\|^2$$

Also, if $f(w)$ is strongly convex $f(w) - \frac{\lambda}{2}\|w\|^2$ is convex

Def.:

A function $f(w)$ is β-Lipschitz on $\mathscr{W}$ if

$$\|\nabla f(w) - \nabla f(w')\| \leq \beta \cdot \|w - w'\|, \forall w, w' \in \mathscr{W}$$

- A quadratic upper bound $f(w) \leq f(w^*) + \frac{\beta}{2}\|w - w^*\|^2$

Str. Convexity & smoothness = co-coercivity

Lemma:

A $f(w)$ that is both λ-strongly and convex on $\mathscr{W}$ if

$$\langle \nabla f(w) - \nabla f(w'), w - w' \rangle \geq \frac{\lambda\beta}{\beta+\lambda}\|w - w'\|^2 + \frac{1}{\beta+\lambda}\|\nabla f(w) - \nabla f(w')\|^2$$

- what does that imply?

$$\langle \nabla f(w), w - w^* \rangle \geq \frac{\lambda\beta}{\beta+\lambda}\|w - w^*\|^2 + \frac{1}{\beta+\lambda}\|\nabla f(w)\|^2$$

Co-coercivity tells us that there is a strong correlation between the gradient of a function and the direction towards optimum, i.e., $\nabla f(w)^T(w - w^*) \geq c_1\|w - w^*\|^2 + c_2\|\nabla f(w)\|^2$

Str. CVX+smoothness = exp. fast convergence

Theorem

Let $f(w)$ be a λ-strongly convex and β-smooth function on $\mathscr{W}$. Then, the iterates of GD with step-size $\gamma = \dfrac{2}{\lambda + \beta}$ satisfy

$$\|w_t - w^*\|^2 \leq e^{-\frac{2t}{\kappa}} \|w_0 - w^*\|^2$$

where $\kappa = \dfrac{\beta}{\lambda}$ is the condition number of $f(w)$.

Str. CVX+smoothness = exp. fast convergence

Theorem

Let $f(w)$ be a λ-strongly convex and β-smooth function on $\mathscr{W}$. Then, the iterates of GD with step-size $\gamma = \dfrac{2}{\lambda + \beta}$ satisfy

$$\|w_t - w^*\|^2 \leq e^{-\frac{2t}{\kappa}} \|w_0 - w^*\|^2$$

where $\kappa = \dfrac{\beta}{\lambda}$ is the condition number of $f(w)$.

Exponentially faster than Lip+cvx!!!

Str. CVX+smoothness = exp. fast convergence

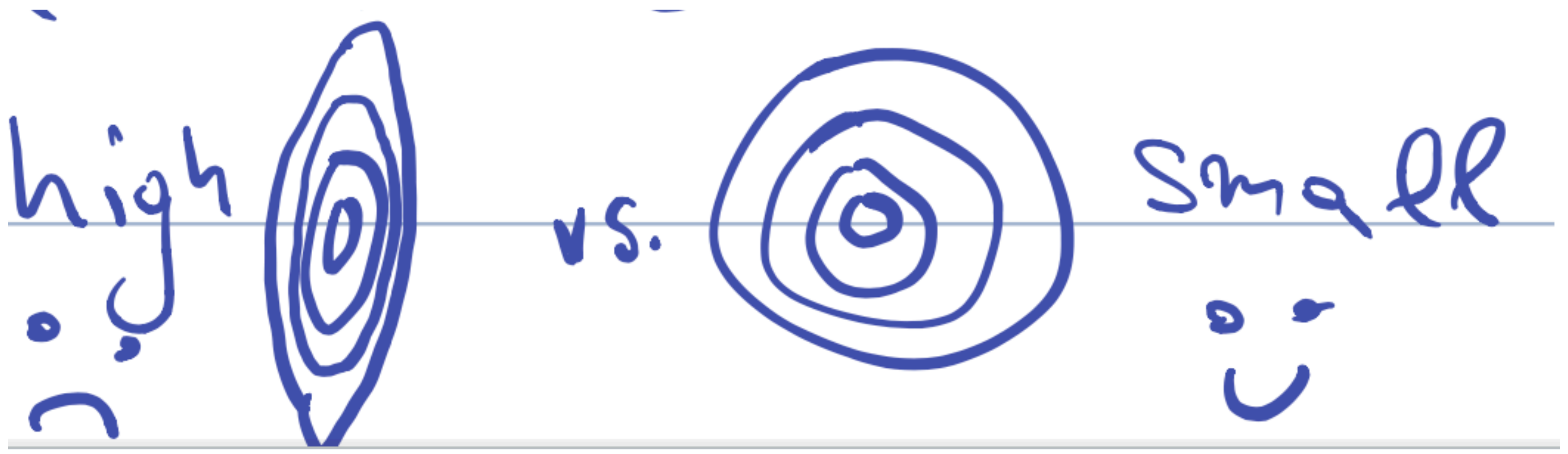
Theorem

Let $f(w)$ be a λ-strongly convex and β-smooth function on $\mathscr{W}$. Then, the iterates of GD with step-size $\gamma = \dfrac{2}{\lambda + \beta}$ satisfy

$$\|w_t - w^*\|^2 \leq e^{-\frac{2t}{\kappa}} \|w_0 - w^*\|^2$$

where $\kappa = \dfrac{\beta}{\lambda}$ is the condition number of $f(w)$.

Exponentially faster than Lip+cvx!!!



Str. CVX+smoothness = exp. fast convergence

Theorem

Let $f(w)$ be a λ-strongly convex and β-smooth function on $\mathscr{W}$. Then, the iterates of GD with step-size $\gamma = \frac{2}{\lambda + \beta}$ satisfy

$$\|w_t - w^*\|^2 \leq e^{-\frac{2t}{\kappa}} \|w_0 - w^*\|^2$$

where $\kappa = \frac{\beta}{\lambda}$ is the condition number of $f(w)$.

- We can get to error ϵ in

$$\epsilon = e^{-\frac{2T}{\kappa}} \|w_0 - w^*\|^2$$

$$\Rightarrow \frac{\epsilon}{\|w_0 - w^*\|^2} = e^{-\frac{2T}{\kappa}}$$

Str. CVX+smoothness = exp. fast convergence

Theorem

Let $f(w)$ be a λ-strongly convex and β-smooth function on $\mathscr{W}$. Then, the iterates of GD with step-size $\gamma = \frac{2}{\lambda + \beta}$ satisfy

$$\|w_t - w^*\|^2 \leq e^{-\frac{2t}{\kappa}} \|w_0 - w^*\|^2$$

where $\kappa = \frac{\beta}{\lambda}$ is the condition number of $f(w)$.

- We can get to error ϵ in

$$\epsilon = e^{-\frac{2T}{\kappa}} \|w_0 - w^*\|^2$$

$$\Rightarrow \frac{\epsilon}{\|w_0 - w^*\|^2} = e^{-\frac{2T}{\kappa}}$$

$$\Rightarrow \ln\left(\frac{\epsilon}{\|w_0 - w^*\|^2}\right) = -\frac{2T}{\kappa}$$

Str. CVX+smoothness = exp. fast convergence

Theorem

Let $f(w)$ be a λ-strongly convex and β-smooth function on $\mathscr{W}$. Then, the iterates of GD with step-size $\gamma = \frac{2}{\lambda + \beta}$ satisfy

$$\|w_t - w^*\|^2 \leq e^{-\frac{2t}{\kappa}} \|w_0 - w^*\|^2$$

where $\kappa = \frac{\beta}{\lambda}$ is the condition number of $f(w)$.

- We can get to error ϵ in

$$\epsilon = e^{-\frac{2T}{\kappa}} \|w_0 - w^*\|^2$$

$$\Rightarrow \frac{\epsilon}{\|w_0 - w^*\|^2} = e^{-\frac{2T}{\kappa}}$$

$$\Rightarrow \ln\left(\frac{\epsilon}{\|w_0 - w^*\|^2}\right) = -\frac{2T}{\kappa}$$

$$T = \frac{\kappa}{2} \ln\left(\frac{\|w_0 - w^*\|^2}{\epsilon}\right)$$

Proof of convergence: Str.CVX+Smooth

- Let us define the iterate difference as $\Delta_T = \|w_{T+1} - w^*\|^2$

$$\|w_{t+1} - w^*\|^2 = \|w_t - \gamma \nabla f(w_t) - w^*\|^2$$

$$\Delta_{t+1} = \Delta_t - 2\gamma\langle \nabla f(w_t), w_t - w^*\rangle + \gamma^2\|\nabla f(w_t)\|^2$$

Proof of convergence: Str.CVX+Smooth

- Let us define the iterate difference as $\Delta_T = \|w_{T+1} - w^*\|^2$

$$\begin{aligned}
\|w_{t+1} - w^*\|^2 &= \|w_t - \gamma \nabla f(w_t) - w * \|^2 \\
\Delta_{t+1} &= \Delta_t - 2\gamma \langle \nabla f(w_t), w_t - w^* \rangle + \gamma^2 \|\nabla f(w_t)\|^2 \\
&\le \Delta_t - 2\gamma \left(\frac{\lambda\beta}{\lambda + \beta} \|w_t - w^*\|^2 + \gamma^2 \|\nabla f(w_t)\|^2 \right)
\end{aligned}$$

Proof of convergence: Str.CVX+Smooth

- Let us define the iterate difference as $\Delta_T = \|w_{T+1} - w^*\|^2$

$$\begin{aligned}
\|w_{t+1} - w^*\|^2 &= \|w_t - \gamma \nabla f(w_t) - w * \|^2 \\
\Delta_{t+1} &= \Delta_t - 2\gamma \langle \nabla f(w_t), w_t - w^* \rangle + \gamma^2 \|\nabla f(w_t)\|^2 \\
&\leq \Delta_t - 2\gamma \left(\frac{\lambda\beta}{\lambda + \beta} \|w_t - w^*\|^2 + \gamma^2 \|\nabla f(w_t)\|^2 \right) \\
&= \|w_t - w^*\| - \frac{4\lambda\beta}{(\lambda+\beta)^2} \Delta_t = \left(1 - \frac{4\lambda\beta}{(\lambda+\beta)^2}\right) \Delta_t
\end{aligned}$$

Proof of convergence: Str.CVX+Smooth

- Let us define the iterate difference as $\Delta_T = \|w_{T+1} - w^*\|^2$

$$
\begin{aligned}
\|w_{t+1} - w^*\|^2 &= \|w_t - \gamma \nabla f(w_t) - w * \|^2 \\
\Delta_{t+1} &= \Delta_t - 2\gamma \langle \nabla f(w_t), w_t - w^* \rangle + \gamma^2 \|\nabla f(w_t)\|^2 \\
&\leq \Delta_t - 2\gamma \Big(\frac{\lambda\beta}{\lambda + \beta} \|w_t - w^*\|^2 + \gamma^2 \|\nabla f(w_t)\|^2 \Big) \\
&= \|w_t - w^*\| - \frac{4\lambda\beta}{(\lambda + \beta)^2} \Delta_t = \Big(1 - \frac{4\lambda\beta}{(\lambda + \beta)^2} \Big) \Delta_t \\
&= \Big(\frac{\lambda^2 + 2\lambda\beta + \beta^2 - 4\lambda\beta}{(\lambda + \beta)^2} \Big) \Delta_t
\end{aligned}
$$

Proof of convergence: Str.CVX+Smooth

- Let us define the iterate difference as $\Delta_T = \|w_{T+1} - w^*\|^2$

$$\|w_{t+1} - w^*\|^2 = \|w_t - \gamma \nabla f(w_t) - w^*\|^2$$

$$\Delta_{t+1} = \Delta_t - 2\gamma \langle \nabla f(w_t), w_t - w^* \rangle + \gamma^2 \|\nabla f(w_t)\|^2$$

$$\leq \Delta_t - 2\gamma \left(\frac{\lambda\beta}{\lambda+\beta} \|w_t - w^*\|^2 + \gamma^2 \|\nabla f(w_t)\|^2 \right)$$

$$= \|w_t - w^*\| - \frac{4\lambda\beta}{(\lambda+\beta)^2} \Delta_t = \left(1 - \frac{4\lambda\beta}{(\lambda+\beta)^2}\right) \Delta_t$$

$$= \left(\frac{\lambda^2 + 2\lambda\beta + \beta^2 - 4\lambda\beta}{(\lambda+\beta)^2} \right) \Delta_t$$

$$\leq \left(\frac{\lambda - \beta}{\lambda+\beta} \right)^2 \Delta_t = \left(\frac{1 - \beta/\lambda}{1 + \beta/\lambda} \right)^2 \Delta_t = \left(\frac{1-\kappa}{1+\kappa} \right)^2 \Delta_t$$

Proof of convergence: Str.CVX+Smooth

- Let us define the iterate difference as $\Delta_T = \|w_{T+1} - w^*\|^2$

$$
\begin{aligned}
\|w_{t+1} - w^*\|^2 &= \|w_t - \gamma \nabla f(w_t) - w * \|^2 \\
\Delta_{t+1} &= \Delta_t - 2\gamma \langle \nabla f(w_t), w_t - w^* \rangle + \gamma^2 \|\nabla f(w_t)\|^2 \\
&\leq \Delta_t - 2\gamma \left(\frac{\lambda\beta}{\lambda + \beta} \|w_t - w^*\|^2 + \gamma^2 \|\nabla f(w_t)\|^2 \right) \\
&= \|w_t - w^*\| - \frac{4\lambda\beta}{(\lambda+\beta)^2} \Delta_t = \left(1 - \frac{4\lambda\beta}{(\lambda+\beta)^2} \right) \Delta_t \\
&= \left(\frac{\lambda^2 + 2\lambda\beta + \beta^2 - 4\lambda\beta}{(\lambda+\beta)^2} \right) \Delta_t \\
&\leq \left(\frac{\lambda - \beta}{\lambda + \beta} \right)^2 \Delta_t = \left(\frac{1 - \beta/\lambda}{1 + \beta/\lambda} \right)^2 \Delta_t = \left(\frac{1-\kappa}{1+\kappa} \right)^2 \Delta_t \\
&\vdots \\
&\leq \left(\frac{1-\kappa}{1+\kappa} \right)^t \cdot \Delta_0 = e^{2t \log(1 - \frac{2}{\kappa})} \cdot \Delta_0
\end{aligned}
$$

Proof of convergence: Str.CVX+Smooth

- Let us define the iterate difference as $\Delta_T = \|w_{T+1} - w^*\|^2$

$$
\begin{aligned}
\|w_{t+1} - w^*\|^2 &= \|w_t - \gamma \nabla f(w_t) - w * \|^2 \\
\Delta_{t+1} &= \Delta_t - 2\gamma\langle \nabla f(w_t), w_t - w^* \rangle + \gamma^2 \|\nabla f(w_t)\|^2 \\
&\leq \Delta_t - 2\gamma\left(\frac{\lambda\beta}{\lambda+\beta}\|w_t - w^*\|^2 + \gamma^2\|\nabla f(w_t)\|^2\right) \\
&= \|w_t - w^*\| - \frac{4\lambda\beta}{(\lambda+\beta)^2}\Delta_t = \left(1 - \frac{4\lambda\beta}{(\lambda+\beta)^2}\right)\Delta_t \\
&= \left(\frac{\lambda^2 + 2\lambda\beta + \beta^2 - 4\lambda\beta}{(\lambda+\beta)^2}\right)\Delta_t \\
&\leq \left(\frac{\lambda-\beta}{\lambda+\beta}\right)^2\Delta_t = \left(\frac{1-\beta/\lambda}{1+\beta/\lambda}\right)^2\Delta_t = \left(\frac{1-\kappa}{1+\kappa}\right)^2\Delta_t \\
&\vdots \\
&\leq \left(\frac{1-\kappa}{1+\kappa}\right)^t \cdot \Delta_0 = e^{2t\log(1-\frac{2}{\kappa})} \cdot \Delta_0 \\
&\leq e^{-\frac{2t}{\kappa}} \cdot \Delta_0
\end{aligned}
$$

Comparison of Convergence Rates

Function Class	Convergence Rate
Lipschitz	$\frac{RL}{\sqrt{T}}$
smooth	$\frac{R^2\beta}{T}$
Lipschitz + str. cvx	$\frac{L^2}{\lambda T}$
smooth + str. cvx	$R^2 e^{-\frac{T}{\kappa}}$

- The structure of a function can help in improving computational complexity. However, we should be cautious that the bounds of complexity are not always tight.

Q: what of these properties are satisfied by practically relevant functions?

Next Time:

Complexity of GD on some practical problems

&

intro to SGD, the simplest learning algorithm

reading list

Bubeck, S., 2015. Convex Optimization: Algorithms and Complexity. Foundations and Trends® in Machine Learning, 8(3-4), pp.231-357.
https://arxiv.org/pdf/1405.4980.pdf

Understanding Machine Learning: From Theory to Algorithms, https://www.cs.huji.ac.il/w~shais/UnderstandingMachineLearning/copy.html