535520: Optimization Algorithms
Lecture 11 — Mirror Descent and Newton's Method

Ping-Chun Hsieh (謝秉均)

November 25, 2024

Announcement

- Lecture today (11/25): 12:20pm-2:20pm
- 10min extension for the next three lectures

This Lecture

1. Mirror Descent

2. Newton's Method

- Reading Material:
  - Amir Beck and Marc Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," Operations Research Letters, 2003.
  - Chapters 3 and 6 of Nocedal & Wright's textbook "Numerical Optimization"
  - Chapter 9 of Stephen Boyd's textbook "Convex Optimization"

Scipy.Optimize in Python?

Optimization and root finding (`scipy.optimize`)

SciPy `optimize` provides functions for minimizing (or maximizing) objective functions, possibly subject to constraints. It includes solvers for nonlinear problems (with support for both local and global optimization algorithms), linear programing, constrained and nonlinear least-squares, root finding, and curve fitting.

Common functions and objects, shared across different solvers, are:

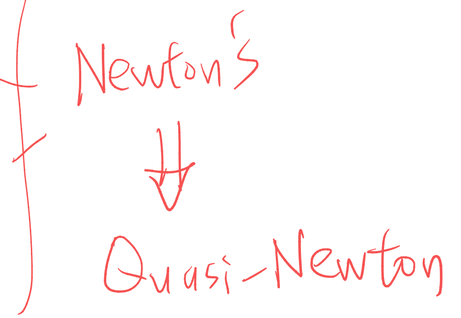
`show_options`([solver, method, disp])	Show documentation for additional options of optimization solvers.
`OptimizeResult`	Represents the optimization result.
`OptimizeWarning`	

Local (multivariate) optimization

`minimize`(fun, x0[, args, method, jac, hess, ...])	Minimization of scalar function of one or more variables.

The **`minimize`** function supports the following methods:

- minimize(method='Nelder-Mead')
- minimize(method='Powell')
- minimize(method='CG')
- minimize(method='BFGS')
- minimize(method='Newton-CG')
- minimize(method='L-BFGS-B')
- minimize(method='TNC')
- minimize(method='COBYLA')



Review: Mirror Descent (MD)

Key Idea: Generalize the L_2 proximal term to other distance measures!

$$x_{t+1} = \arg\min_{x \in C} \left\{ f(x_t) + \nabla f(x_t)^\top (x - x_t) + \frac{1}{\eta_t} D_\phi(x \| x_t) \right\}$$

first-order approximation

Bregman divergence

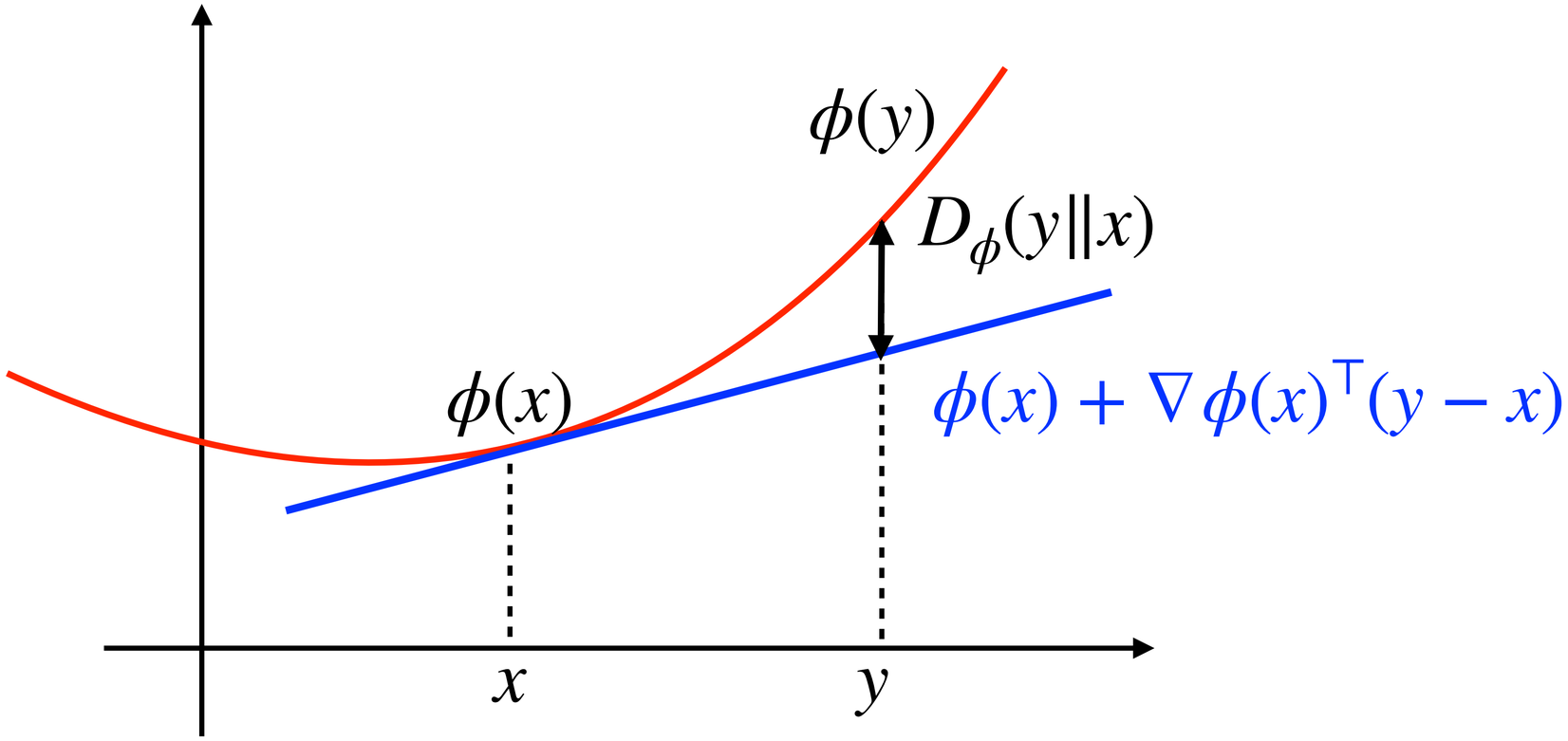
where $D_\phi(y \| x) := \phi(y) - \phi(x) - \nabla \phi(x)^\top (y - x)$

(with respect to $\phi(\cdot)$ strictly convex and differentiable)

- Remark: Bregman divergence is meant to capture "local geometry" of objective function
- Remark: Bregman divergence is NOT symmetric and hence not a metric
- How about MD in the unconstrained cases?

Review: Bregman Divergence

$$D_\phi(y\|x) = \phi(y) - \left(\phi(x) + \nabla\phi(x)^\top(y-x)\right),$$ where $\phi(\,\cdot\,)$ is strictly convex



Review: Basic Properties of Bregman Divergence

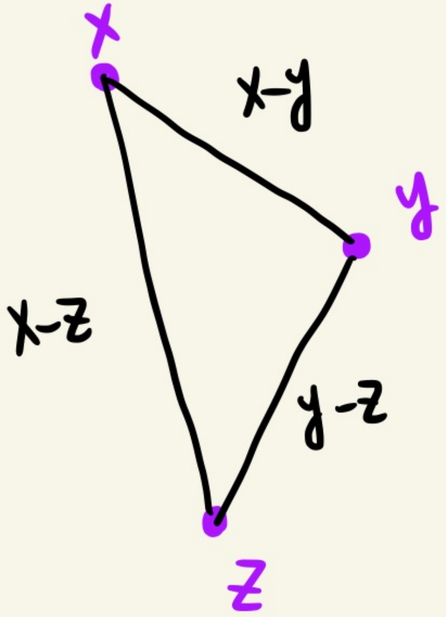
Let $\phi : X \to \mathbb{R}$ be a strictly convex and differentiable function

- 1. Non-negativity: $D_\phi(y\|x) \geq 0$
- 2. Distance between a point to itself is zero: $D_\phi(y\|x) = 0$ if and only if $x = y$
- 3. Convexity: $D_\phi(y\|x)$ is convex in y (but not necessarily in x)
- 4. Symmetry not guaranteed: $D_\phi(y\|x) \neq D_\phi(x\|y)$ in general

Bregman Divergence: Three-Point Lemma

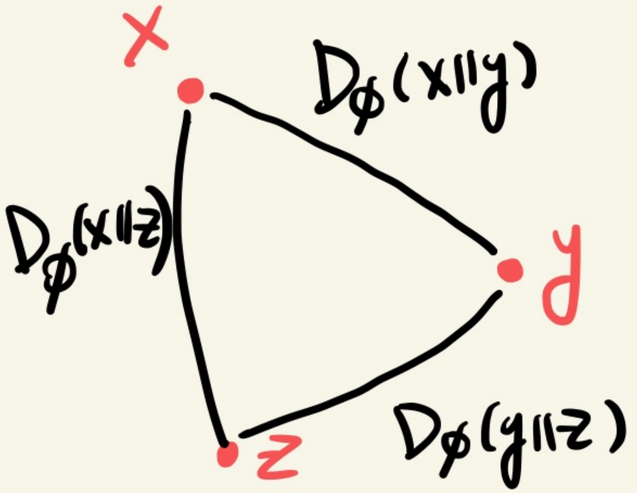
Euclidean distance

$$\|x-z\|^2 = \|x-y\|^2 + \|y-z\|^2 - 2(z-y)^T(x-y)$$



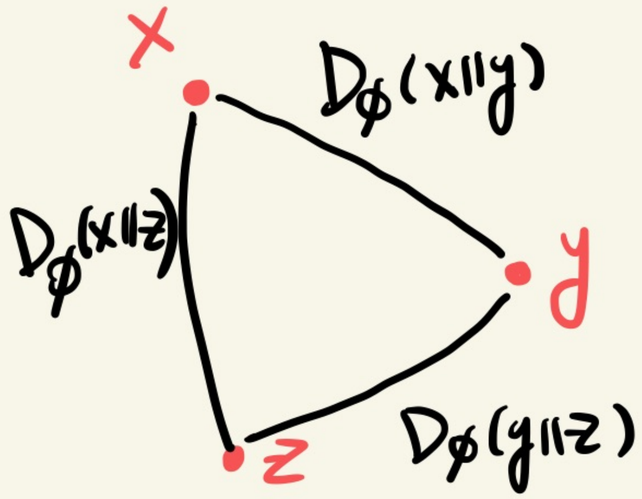
Bregman divergence

$$D_\phi(x\|z) = D_\phi(x\|y) + D_\phi(y\|z) - \left(\nabla\phi(z) - \nabla\phi(y)\right)^T(x-y)$$



Proof of Three-Point Lemma

$$
\begin{aligned}
& D_\phi(x\|y) + D_\phi(y\|z) - D_\phi(x\|z) \\
= & \Big(\phi(x) - \phi(y) - \nabla\phi(y)^T(x-y) \Big) \\
& + \Big(\phi(y) - \phi(z) - \nabla\phi(z)^T(y-z) \Big) \\
& - \Big(\phi(x) - \phi(z) - \nabla\phi(z)^T(x-z) \Big) \\
= & \big(\nabla\phi(z) - \nabla\phi(y) \big)^T (x-y)
\end{aligned}
$$



Another Viewpoint: MD as *Follow-The-Regularized-Leader* in Online Linear Optimization

Chapter 28 of Csaba Sepesvari's textbook "Bandit Algorithms"

A Motivating Example of Online Linear Optimization

- Example: **Expert Problem**
  - A learner and n experts
  - At each time t, the learner chooses an action probability vector $a_t \in \Delta_n$ (indicating the distribution of expert-following behavior of the learner)
  - Let $y_t \in \mathbb{R}_+^n$ denote the cost vector of following the experts (known after choosing a_t)
  - Goal: Minimize "total cost" = Minimize "regret" (relative to a fixed comparator $a \in \Delta_n$)

$$R_T := \max_{a \in \Delta_n} \left\{ \sum_{t=1}^{T} (a_t - a)^\top y_t \right\}$$

Suppose $n=3$ $a_t = \begin{bmatrix} 0.2 \\ 0.5 \\ 0.3 \end{bmatrix}$ "adversarial"

$\begin{bmatrix} 2.5 \\ 5 \\ 10 \end{bmatrix}$

$\min \sum_{t=1}^{T} a_t^\top y_t$

At each time t:

Based on the history $\{y_1, y_2, \cdots y_{t-1}\}$, select your $a_t \in \Delta_n$

Interesting Fact: $R_T = O(\sqrt{T \log n})$ under properly-designed algorithms!

MD and "Follow-The-Regularized-Leader" (FTRL)

Mirror Descent

Initially:

$$a_1 = \arg\min_{a \in A} \phi(a)$$

For $t \geq 1$:

$$a_{t+1} = \arg\min_{a \in A} \{ a^\top y_t + \frac{1}{\eta} D_\phi(a \| a_t) \}$$

fine-tuning based on the new observation

fine-tuning based on the current action a_t

Follow-The-Regularized-Leader

Initially:

$$a_1 = \arg\min_{a \in A} \phi(a)$$

For $t \geq 1$:

$$a_{t+1} = \arg\min_{a \in A} \{ \sum_{s=1}^{t} a^\top y_s + \frac{1}{\eta} \phi(a) \}$$

(Choosing the action that appears the best in hindsight under regularization)

Interesting Fact: MD and FTRL are equivalent! (under some mild conditions)

Equivalence Between MD and FTRL

MD $a_{t+1} = \underset{a \in A}{\operatorname{argmin}} \left\{ a^T y_t + \frac{1}{\eta} \cdot D_\phi(a \| a_t) \right\}$ $\quad = F(a)$

$$F(a) = a^T y_t + \frac{1}{\eta}\left(\phi(a) - \phi(a_t) - \nabla\phi(a_t)^T(a - a_t)\right)$$

$$\Rightarrow \nabla F(a) = y_t + \frac{1}{\eta}\left(\nabla\phi(a) - \nabla\phi(a_t)\right)$$

If a_{t+1} is in the interior of A, then

By FONC: $\nabla F(a_{t+1}) = 0$ (and hence $-\eta \cdot y_t = \nabla\phi(a_{t+1}) - \nabla\phi(a_t)$)

Therefore, by taking "telescoping sum", we have

$$\nabla\phi(a_{t+1}) = -\eta \cdot \sum_{s=1}^{t} y_t$$

$$\Rightarrow a_{t+1} = (\nabla\phi)^{-1}\left(-\eta \cdot \sum_{s=1}^{t} y_t\right)$$

Equivalence Between MD and FTRL

FTRL $a_{t+1} = \operatorname{argmin}_{a \in A} \left\{ \underbrace{\sum_{s=1}^{t} a^\top y_s + \frac{1}{\eta}\phi(a)}_{G(a)} \right\}$

If a_{t+1} is in the interior of A, then $\nabla G(a_{t+1}) = 0$ (by FONC)

(Equivalently, $\nabla \phi(a_{t+1}) = -\eta \cdot \sum_{s=1}^{t} y_s$)

Why Not Using "Follow-The-Leader" (FTL)?

Follow-The-Leader: Choosing the action that appears the best in hindsight

$$a_{t+1} = \arg\min_{a \in A} \left\{ \sum_{s=1}^{t} a^\top y_s \right\}$$

- FTL (i.e., without a regularizer) suffers from linear regret!

- Counterexample: action: 1-dimensional
  - Let action set $A = [-1,1]$ and initial action $a_1 = 0$
  - Cost vectors: $y_1 = 1/2$ and $y_s = (-1)^{s+1}, \forall s > 1$

Counterexample of "Follow-The-Leader"

$A = [-1, 1]$, $a_1 = 0$

Total cost up to time t: $\sum_{s=1}^{t} a_s^T y_s = t - 1$

$y_1 = \frac{1}{2} \Rightarrow$ $a_2 = \underset{a \in A}{\operatorname{argmin}}\ a^T y_1 =$ -1

$y_2 = -1 \Rightarrow$ $a_3 = \underset{a \in A}{\operatorname{argmin}}\ a^T (y_1 + y_2) =$ $+1$ ($y_1 + y_2 = -\frac{1}{2}$)

$y_3 = +1 \Rightarrow$ $a_4 = \underset{a \in A}{\operatorname{argmin}}\ a^T (y_1 + y_2 + y_3) =$ -1 ($y_1 + y_2 + y_3 = \frac{1}{2}$)

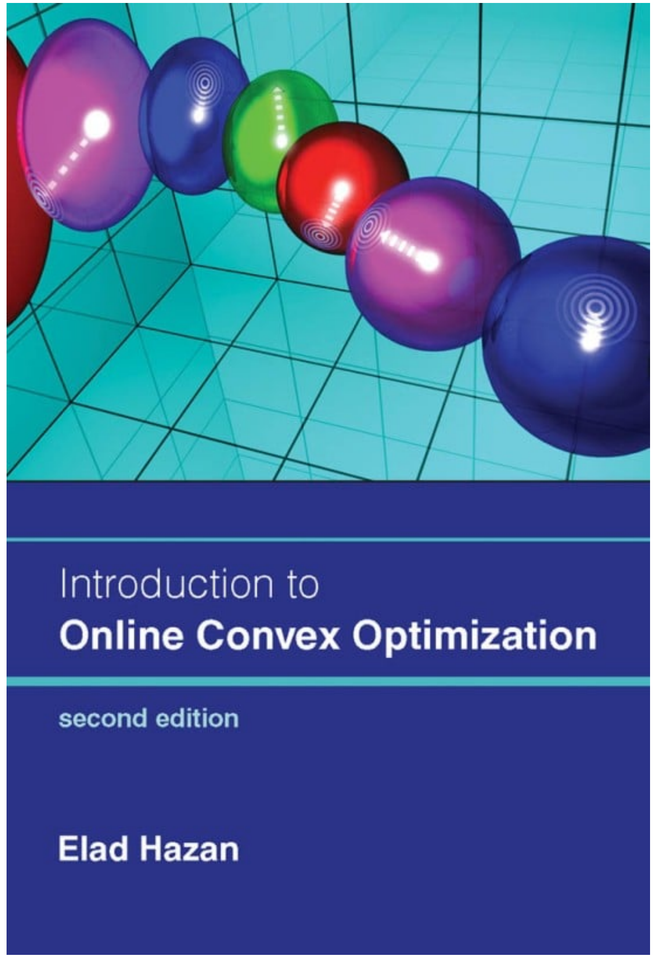
$\vdots$

Hence, $a_t =$ $(-1)^{t+1}$

"Best action in hindsight" is 0, Regret $R_T =$ $T - 1$

We know $y_1, \cdots, y_t$

https://tor-lattimore.com/downloads/book/book.pdf



https://arxiv.org/abs/1909.05207

Convergence of Mirror Descent

Convergence of MD Algorithms

$\frac{1}{p} + \frac{1}{q} = 1.$

Theorem Suppose the following conditions hold:

① f is Convex and L_f-Lipschitz continuous (in the sense that $\|\nabla f(x)\|_q \le L_f$)

② ϕ is ρ-Strongly convex w.r.t. $\|\cdot\|_p$

Then, MD achieves

$$f_{\text{best}}^t - f^* \le \underbrace{\sup_{x \in C} D_\phi(x \| x_0)}_{=: D_{\max}} + \frac{L_f^2}{2\rho} \cdot \sum_{\tau=0}^{t} \eta_\tau^2 \Big/ \sum_{\tau=0}^{t} \eta_\tau$$

(η_τ: step size)

Moreover, by letting $\eta_t = \frac{\sqrt{2\rho \cdot D_{\max}}}{L_f} \cdot \frac{1}{\sqrt{t}}$, $= O\left(\frac{1}{\sqrt{t}}\right)$

$$f_{\text{best}}^t - f^* = O\left(\frac{L_f \sqrt{D_{\max}}}{\sqrt{\rho}} \cdot \frac{\log t}{\sqrt{t}} \right)$$

Amir Beck and Marc Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” Operations Research Letters, 2003.

Example: Optimization over Probability Simplex

(For simplicity, let $X_0 = (\frac{1}{d}, \cdots, \frac{1}{d})$)

Suppose we consider: $\min f(x)$, subject to $x \geq 0$, $1^T x = 1$.

Let's compare MD with "Euclidean norm" vs "KL divergence"

---

① Euclidean norm:

- $\phi(x) = \frac{1}{2}\|x\|^2$ is 1-strongly convex w.r.t. $\|\cdot\|_2$

By the definition of Bregman divergence

$$\phi(y) - (\phi(x) + \nabla\phi(x)^T(y-x)) \geq \frac{1}{2}\|y-x\|^2$$

$$D_{max} := \sup_{x \in C} D_\phi(x \| x_0) = \sup_{x \in C} \frac{1}{2}\left\| x - (\tfrac{1}{d}, \cdots, \tfrac{1}{d}) \right\|^2 \leq \frac{1}{2}$$

$$\text{Hence, } f_{best}^t - f^* = O\left(\frac{L_f \cdot \sqrt{D_{max}}}{\sqrt{\rho}} \cdot \frac{\log t}{\sqrt{t}} \right) = O\left(L_{f,2} \cdot \frac{\log t}{\sqrt{t}} \right)$$

$\sqrt{D_{max}} = \frac{1}{2}$

$\rho = 1$

$$g_{ap}(t+1) \leq \underbrace{\text{const}}_{\downarrow} \cdot g_{ap}(t) + \alpha \quad \begin{pmatrix} GD \\ NAG \end{pmatrix}$$

$$g_{ap}(t) = \frac{1}{t}$$

$$g_{ap}(t) = \beta^t$$

$$\beta^{t+1} \leq (<1)\,\beta^t + 0$$

(Cont.)

② KL divergence:

- $\phi(x) = -\sum_{i=1}^{d} x^{(i)} \log x^{(i)}$ is 1-strongly convex w.r.t. $\|\cdot\|_1$

$$\sup_{x \in C} D_\phi(x \| x_0) = \sup_{x \in C} KL(x \| x_0) = \sup_{x \in C} \left(\sum_{i=1}^{d} x^{(i)} \log x^{(i)} - \sum_{i=1}^{d} x^{(i)} \cdot \log \frac{1}{d} \right)$$

$$= \log d + \sup_{x \in C} \sum_{i=1}^{d} x^{(i)} \log x^{(i)} \leq \log d$$

Hence, $f_{best}^t - f^* = O\left(\frac{L_f \sqrt{D_{max}}}{\sqrt{\rho}} \cdot \frac{\log t}{\sqrt{t}} \right) = O\left(L_{f,\infty} \cdot \sqrt{\log d} \cdot \frac{\log t}{\sqrt{t}} \right)$

Comparison:

	Euclidean	KL Divergence
Convergence Rate	$O\left(L_{f,2} \cdot \frac{\log t}{\sqrt{t}}\right)$	$O\left(L_{f,\infty} \cdot \sqrt{\log d} \cdot \frac{\log t}{\sqrt{t}}\right)$

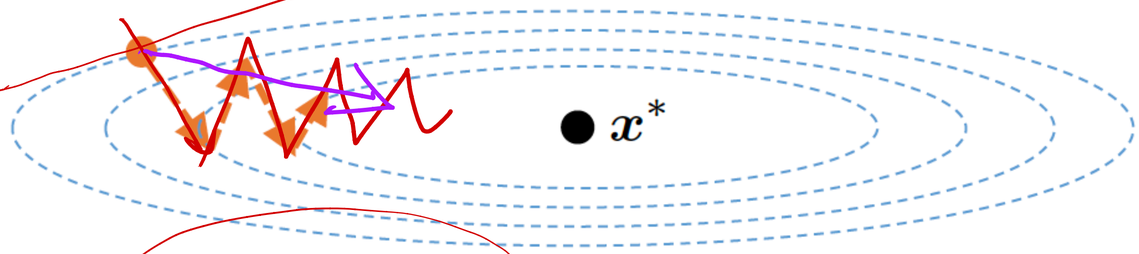
Note that $\|\nabla f\|_\infty \le \|\nabla f\|_2 \le \sqrt{d} \cdot \|\nabla f\|_\infty$ (why?)

Hence, $\frac{1}{\sqrt{d}} \le \frac{L_{f,\infty}}{L_{f,2}} \le 1 \Rightarrow$ MD with KL divergence has a better convergence rate!

Newton’s Method

Recall: Non-Homogeneity and "Scaled" Gradients

$$\text{minimize}_{x \in \mathbb{R}^2} \quad f(x) := \frac{1}{2}(x - x^*)^\top Q (x - x^*)$$



- Suppose $Q = [Q_{11}, 0; 0, Q_{22}]$ is a diagonal matrix with $Q_{11} \gg Q_{22}$

- Idea: Accelerate GD by *scaling the gradient*

$$x_{t+1} = x_t - \eta_t Q^{-1} \nabla f(x_t) = \underline{\qquad\qquad}$$

Notably, in quadratic problems, we also have

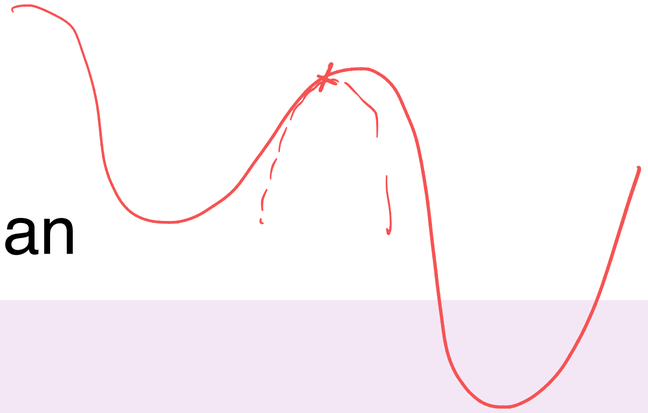
$$\nabla^2 f(x) = \underline{\qquad\qquad}$$

- Question: How about "scaled gradients" for objectives *beyond quadratic functions*?

Newton's Method

For an unconstrained problem: $\text{minimize}_{x \in \mathbb{R}^d} \quad f(x)$

Newton's method = "scaled" gradient by inverse of Hessian



$$x_{t+1} = x_t - \eta_t \underbrace{(\nabla^2 f(x_t))^{-1} \nabla f(x_t)}_{=:\Delta x_{NT}}$$

Newton's step

- Pure Newton's step: Set $\eta_t = 1$
- Damped Newton's step: η_t is determined by "backtracking line search" (discussed later)

- Question: Is Newton's method a "descent" method?

If you choose an update direction z,

$\nabla f(x)^T z \leq 0$

$\nabla f(x_t)^T \left((\nabla^2 f(x_t))^{-1} \nabla f(x_t) \right) \geqslant 0$?

Interpretations of Newton's Step: (1) Minimizer of Second-Order Approximation

Newton's step $\Delta x_{NT} := (\nabla^2 f(x))^{-1} \nabla f(x)$

FONC: $\nabla f(x) = \nabla^2 f(x) \cdot z$

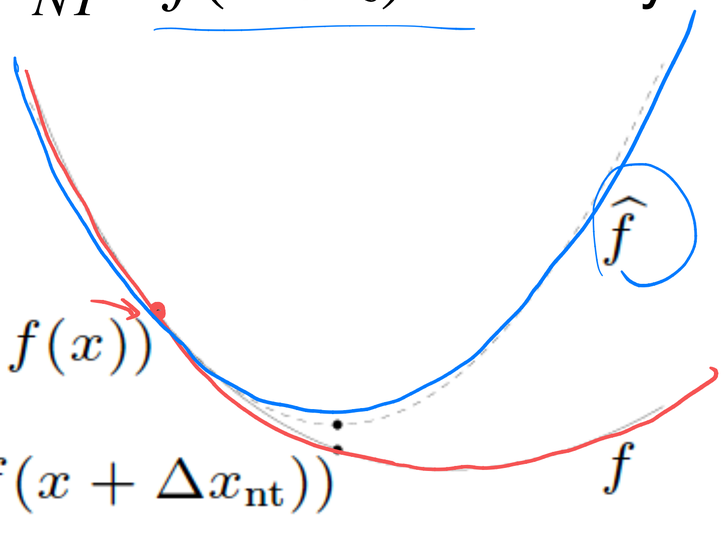
$\Rightarrow z = (\nabla^2 f(x))^{-1} \nabla f(x)$

- Second-order approximation of f at some given point x:

$$\hat{f}(x+z) = f(x) + \nabla f(x)^\top z + \frac{1}{2} z^\top \nabla^2 f(x) z$$

- $\hat{f}(x+z)$, as a function of z, is minimized at $z = \Delta x_{NT}$ if $\hat{f}(x+z)$ is strictly convex in z
- If $\nabla f(x) = 0$, then $\Delta x_{NT} = 0$

Takeaway: Newton's step finds the minimizer of second-order approximation (under $\hat{f}$ strictly convex)



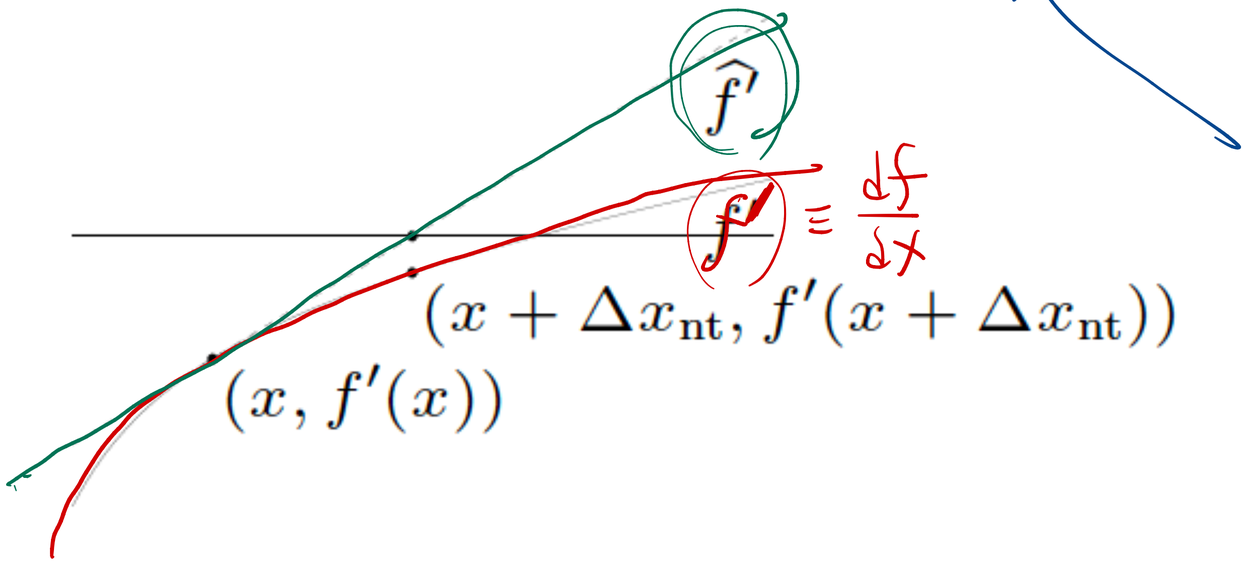
(Figure Source: Chapter 9 of Stephen Boyd's textbook)

Interpretations of Newton's Step: (2) Solution to Linearized Optimality Condition

Newton's step $\Delta x_{NT} := (\nabla^2 f(x))^{-1} \nabla f(x)$

- First-order necessary condition near x under linear approximation

$$\nabla f(x+z) \approx \nabla f(x) + (\nabla^2 f(x)) z = 0 \quad \Rightarrow \quad z = (\nabla^2 f(x))^{-1} \nabla f(x) \equiv \Delta x_{NT}$$



$$\hat{f}(x+z) = f(x) + \nabla f(x)^T z + \frac{1}{2} z^T \nabla^2 f(x) z$$

$$\nabla \hat{f}(x+z) = \nabla f(x) + \left(\nabla^2 f(x)\right) \cdot z$$

(Figure Source: Chapter 9 of Stephen Boyd's textbook)

Features of Newton's Method

1. Invariance under (invertible) linear transformation of coordinates
   - Let $P : \mathbb{R}^d \to \mathbb{R}^d$ be an invertible linear transformation and $x = Py$
   - Then, we have $x + \Delta x_{NT} = P(y + \Delta y_{NT})$

2. "Superlinear" "local" convergence
   - Superlinear: $\|x_{t+1} - x^*\| \leq \gamma \|x_t - x^*\|^\alpha$ with $\alpha > 1$ (x^* is a stationary point)
   - Local: The above holds only if one starts within a small neighborhood of x^*
   - Note: The above does NOT imply convergence to local minimum!

Recall: Terminology of Convergence Rates

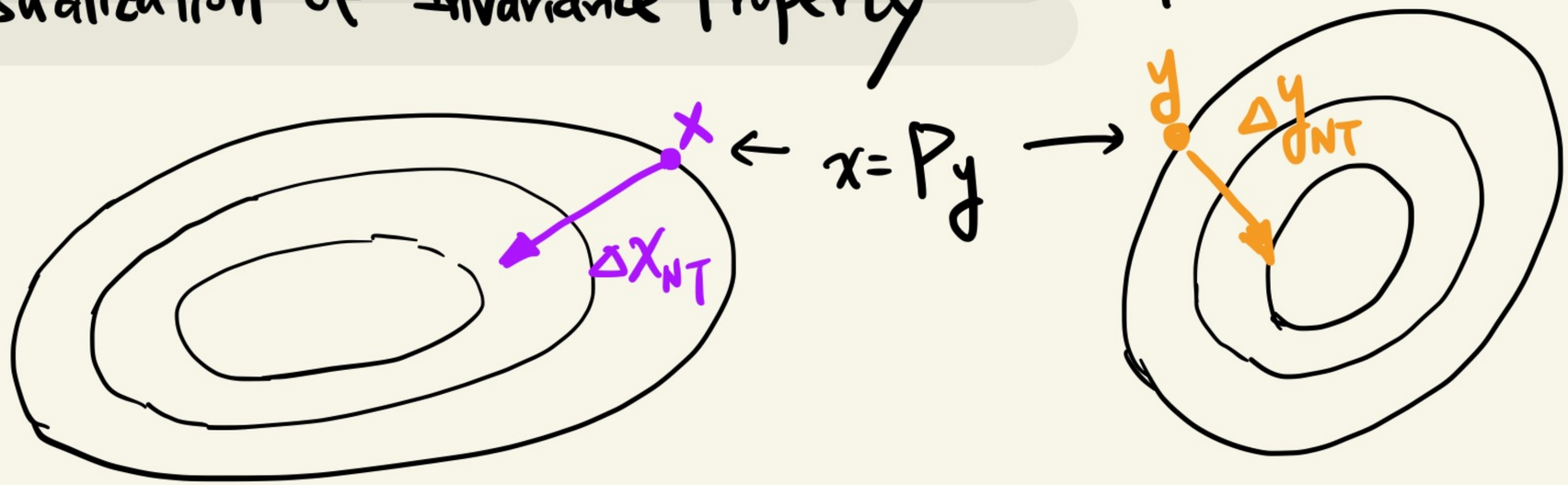
- Let $e(x)$ denote the distance from optimality
  - Example: $e(x) = \|x - x^*\|$
  - Example: $e(x) = |f(x) - f(x^*)|$
- Rate of convergence: The limit of the ratio of successive errors

$$\lim_{k \to \infty} \frac{e(x_{k+1})}{e(x_k)} = \beta$$

  - If $\beta = 1$: We call it a sub-linear rate of convergence
  - If $\beta \in (0,1)$: We call it a linear rate of convergence
  - If $\beta = 0$: We call it a super-linear rate of convergence

Visualization of Invariance Property

(P is an invertible matrix)



- Define $h(y) = f(Py)$

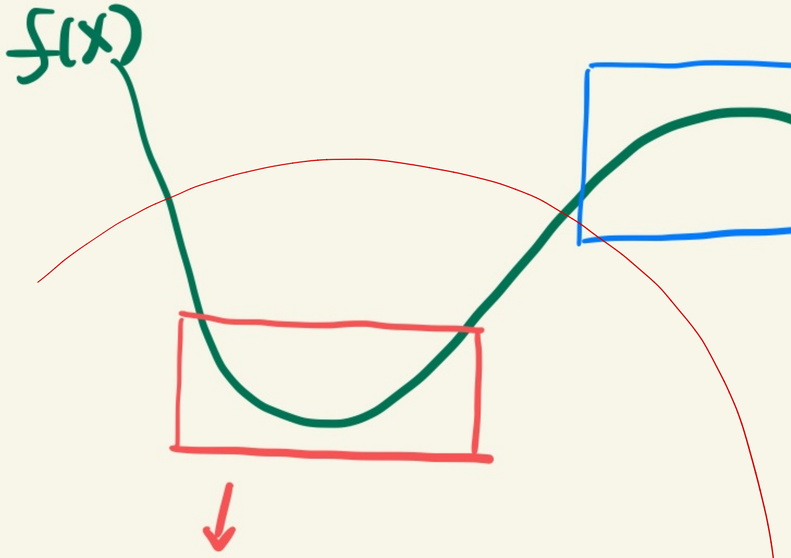
$$\Rightarrow \begin{cases} \nabla h(y) = P^T \nabla f(Py) = P^T \nabla f(x) \\ \nabla^2 h(y) = P^T (\nabla^2 f(x)) P \end{cases}$$

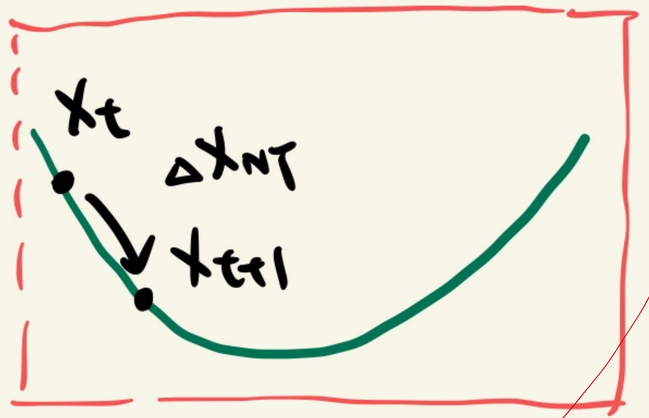
$$\Delta y_{NT} = -(P^T \nabla^2 f(x) P)^{-1} (P^T \nabla f(x))$$

$$= P^{-1} \left(-(\nabla^2 f(x))^{-1} \nabla f(x) \right)$$

Convergence Issue of Pure Newton's Step

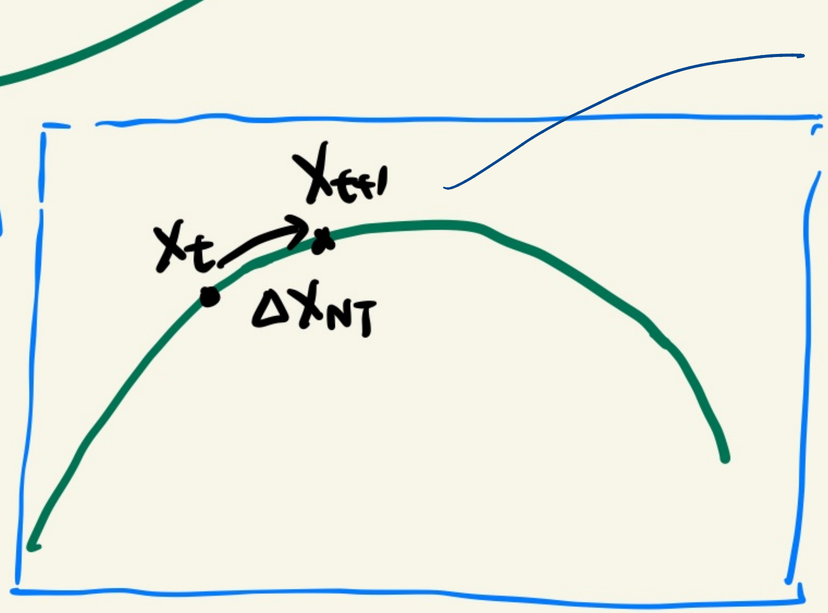
$$\Delta X_{NT} = \left(\nabla^2 f(x)\right)^{-1} \nabla f(x)$$





$\frac{df}{dx} < 0$

$\frac{d^2f}{dx^2} > 0$



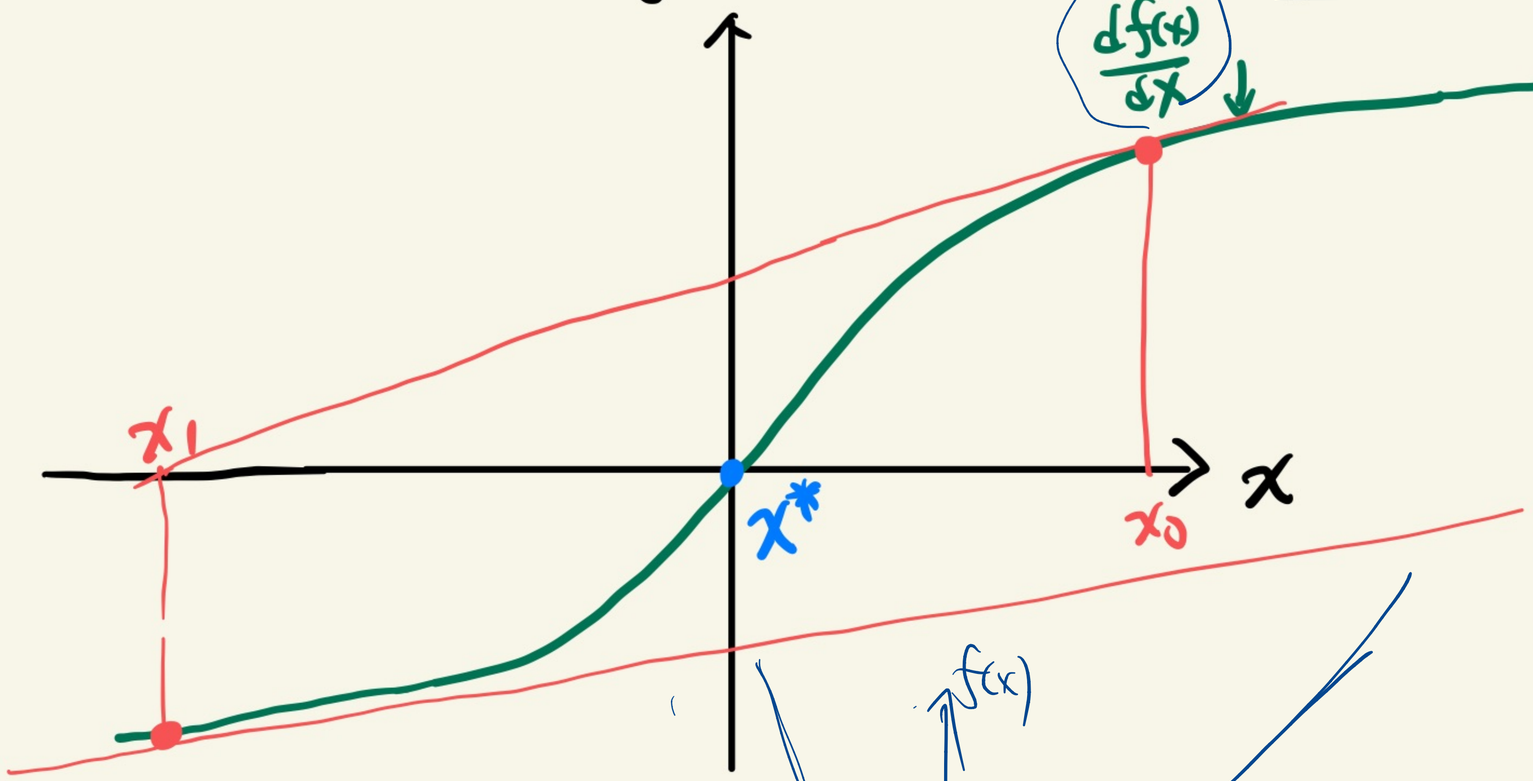
$\frac{df}{dx} > 0$

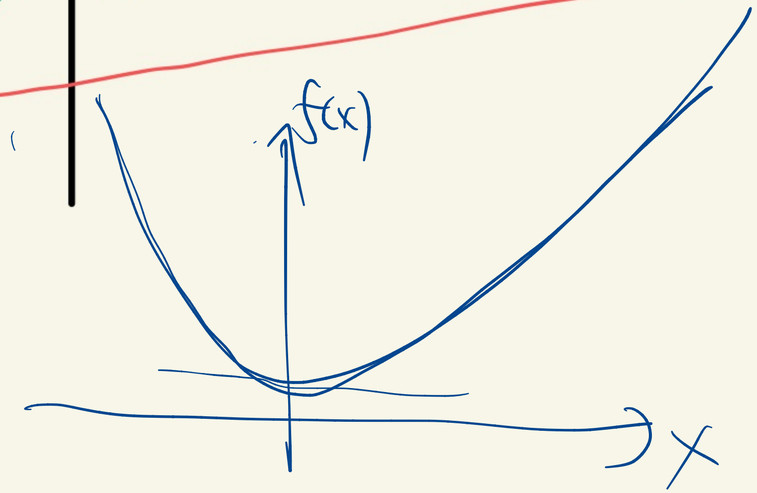
$\frac{d^2f}{dx^2} < 0$

- Pure Newton's Step can get stuck at local maximum!
- This is due to that Pure Newton's step looks for $\nabla f(x) = 0$

Convergence Issue of Pure Newton's Step

Consider a "strongly-convex" function $f(x): \mathbb{R}^1 \to \mathbb{R}^1$ with $\frac{df(x)}{dx}$ as:





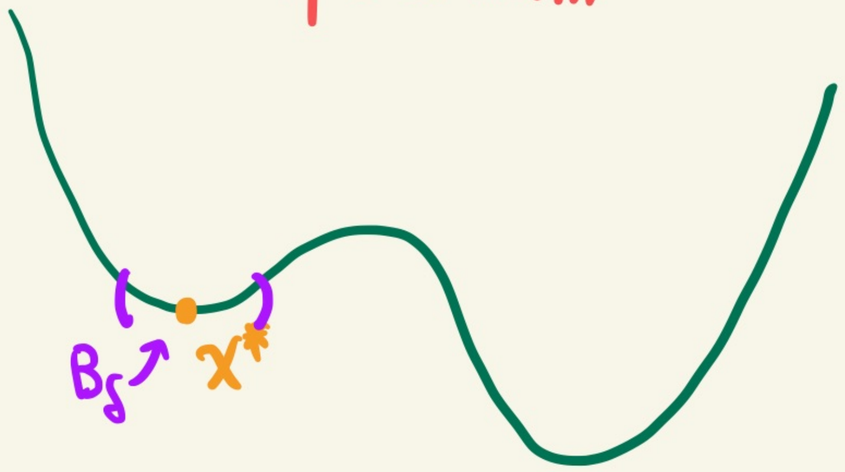
If x_0 is far away from x^*, then pure Newton's step could lead to divergence!

Convergence of Newton's Method: Local Convergence

Theorem Let $f(x)$ be twice continuously differentiable (could be non-convex)

- Let x^* be a stationary point (i.e., $\nabla f(x^*) = 0$) and define $B_\delta := \{x' : \|x' - x^*\| \le \delta\}$
- Hessian of $f(x)$ is L-Lipschitz continuous, i.e., $\|\nabla^2 f(x) - \nabla^2 f(y)\| \le L \cdot \|x - y\|$
- Suppose there exists $\delta > 0$ s.t. $(\nabla^2 f(x))^{-1}$ exists and $\|\nabla^2 f(x)^{-1}\| \le M$, forall $x \in B_\delta$.
  - ↑ spectral norm
- The initial point $x_0 \in B_\delta$.

Then, $$\|x_{t+1} - x^*\| \le \frac{LM}{2} \cdot \|x_t - x^*\|^2$$



A Useful Lemma: Gradient Theorem for Line Integrals

Lemma: Let $g(x): \mathbb{R}^d \to \mathbb{R}$ be a scalar-valued differentiable function.

Then, for any "continuous curve" ϕ which starts at x_1 and ends at x_2

$$\int_{\phi} \nabla g(z)^T dz = g(x_2) - g(x_1)$$

- If ϕ is a line segment from x_1 to x_2:

$$g(x_2) - g(x_1) = \int_{\phi} \nabla g(z)^T dz = \left(\int_0^1 \nabla g\left(x_1 + t \cdot (x_2 - x_1)\right)^T dt \right) (x_2 - x_1)$$

Proof of Convergence

Step 1:

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

$$= \|(\nabla^2 f(x_t))^{-1}\left((\nabla^2 f(x_t))\cdot(x_t - x^*) - \nabla f(x_t)\right)\|$$

$$\nabla f(x_t) = \left(\int_0^1 \nabla^2 f(x^* + s\cdot(x_t - x^*))^T ds\right)\cdot(x_t - x^*)$$

$$= \|(\nabla^2 f(x_t))^{-1}\left(\int_0^1 (\nabla^2 f(x_t)) - \nabla^2 f(x^* - s(x_t - x^*))^T ds\right)\cdot(x_t - x^*)\|$$

$$\leq \|(\nabla^2 f(x_t))^{-1}\| \cdot \|\int_0^1 (\nabla^2 f(x_t)) - \nabla^2 f(x^* - s(x_t - x^*))^T ds\| \cdot \|x_t - x^*\|$$

(Cont.).

Step2: $\left\| \int_0^1 \left(\nabla^2 f(x_t)\right) - \nabla^2 f(x^* - s(x_t - x^*))\, ds \right\|$

Triangle inequality ↓

$$\leq \int_0^1 \left\| \nabla^2 f(x_t) - \nabla^2 f(x^* - s(x_t - x^*)) \right\| ds$$

$\leq L \cdot s \cdot \|x_t - x^*\|$ by smoothness

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

Step3: By combining Step1 and Step2,

$$\|x_{t+1} - x^*\| \leq \frac{LM}{2} \cdot \|x_t - x^*\|^2.$$

□

Remark on Newton's Decrement

$$\lambda(x)^2 := \nabla f(x)^\top (\nabla^2 f(x))^{-1} \nabla f(x) \equiv \nabla f(x)^\top \Delta x_{NT}$$

- Newton's decrement = $f(x) - \inf_z \hat{f}(z)$ (where $\hat{f}(z)$ is the second-order approximation)

$$f(x) - \inf_z \hat{f}(z) = f(x) - \hat{f}(x + \Delta x_{NT}) = \frac{1}{2}\lambda(x)^2$$

- Hence, Newton's decrement is a good estimate of sub-optimality gap $f(x) - f(x^*)$

Summary: Pros and Cons of Newton's Method

$$\text{minimize}_{x \in \mathbb{R}^d} \quad f(x)$$

$$x_{t+1} = x_t - \eta_t \underbrace{(\nabla^2 f(x_t))^{-1} \nabla f(x_t)}_{=:\Delta x_{NT}}$$

- Pros
  - Fast convergence (cf. iteration complexity = $O(\log\log\frac{1}{\epsilon})$)
- Cons
  - Requires storing and inverting Hessian $\nabla^2 f(x) \in R^{d \times d}$
  - One single iteration could take forever
  - Prohibitively large storage overhead

Next Topic: Can we do Newton's step without Hessian?

Quasi-Newton Methods!

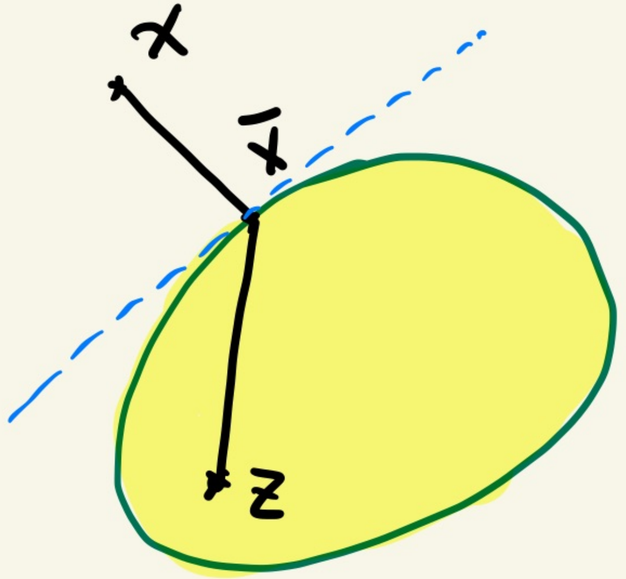
Appendix: Proof of Convergence of Mirror Descent

Amir Beck and Marc Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” Operations Research Letters, 2003.

Generalized Pythagorean Theorem

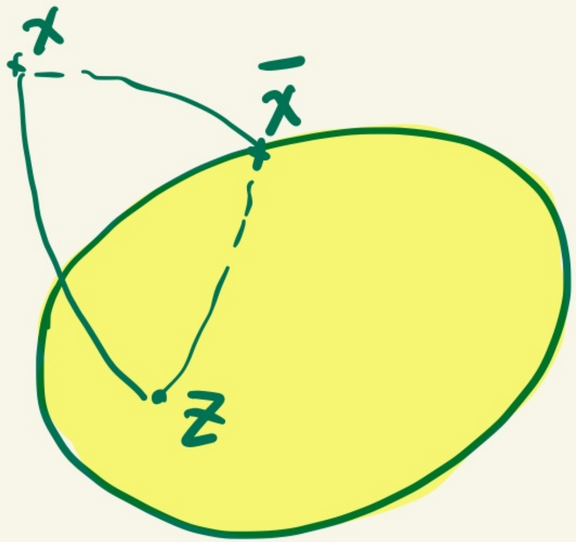
Euclidean Case

$$\|z-x\|^2 \geq \|\bar{x}-z\|^2 + \|x-\bar{x}\|^2$$



Bregman Divergence

$$D_\phi(z\|x) \geq D_\phi(z\|\bar{x}) + D_\phi(\bar{x}\|x).$$



(Proof: HW2)

Hölder's Inequality

Lemma: Let $p, q \in [1, \infty]$ so that $\frac{1}{p} + \frac{1}{q} = 1$. Then, for any two vectors $x = (x_1, \cdots, x_n)$, $y = (y_1, \cdots, y_n)$, we have

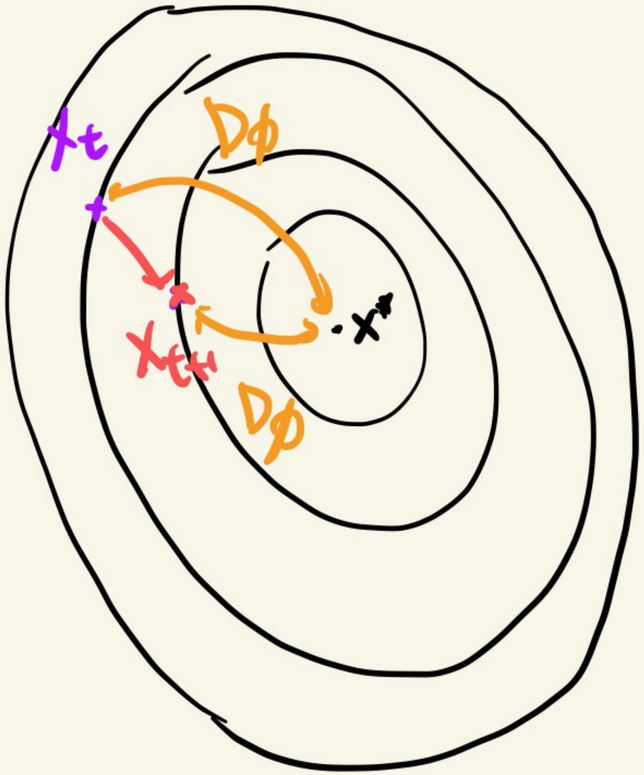
$$\left| \sum_{i=1}^{n} x_i \cdot y_i \right| \leq \|x\|_p \cdot \|y\|_q .$$

Remark: If $p=2, q=2$, Hölder's inequality is essentially Cauchy–Schwarz

A Fundamental Inequality of MD

Lemma $$\eta_t \cdot (f(x_t) - f^*) \le D_\phi(x^* \| x_t) - D_\phi(x^* \| x_{t+1}) + \frac{\eta_t^2 L_f^2}{2\cdot\rho}$$

Question: What's the intuition?



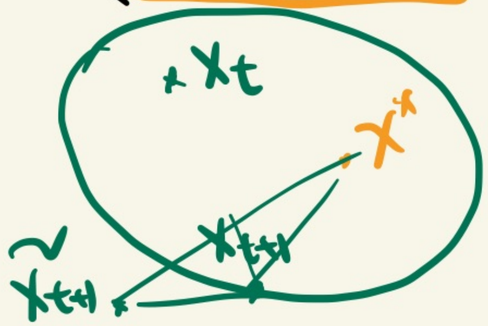
Proof of Lemma:

Step 1: $f(x_t) - f(x^*) \le \nabla f(x_t)^\top (x_t - x^*)$ … (By convexity)

$= \frac{1}{\eta_t} \cdot \left(\nabla \phi(x_t) - \nabla \phi(\tilde{x}_{t+1}) \right)^\top (x_t - x^*)$ … (By the “mirror map”)

$= \frac{1}{\eta_t} \cdot \left(D_\phi(x^* \| x_t) + D_\phi(x_t \| \tilde{x}_{t+1}) - D_\phi(x^* \| \tilde{x}_{t+1}) \right)$ … (By Three-Point Lemma)

$\le \frac{1}{\eta_t} \left(D_\phi(x^* \| x_t) + D_\phi(x_t \| \tilde{x}_{t+1}) - \left(D_\phi(x^* \| x_{t+1}) + D_\phi(x_{t+1} \| \tilde{x}_{t+1}) \right) \right)$ … (By Generalized Pythagorean Theorem)



(Cont.).

Step2: It is sufficient to show the following claim:

$$D_\phi(x_t, \tilde{x}_{t+1}) - D_\phi(x_{t+1}, \tilde{x}_{t+1}) \leq \frac{(\eta_t \cdot L_f)^2}{2\cdot\rho} \quad \text{——(*)}$$

(which would naturally lead to our required lemma)

$$D_\phi(x_t, \tilde{x}_{t+1}) - D_\phi(x_{t+1}, \tilde{x}_{t+1})$$

$$= \phi(x_t) - \phi(x_{t+1}) - \nabla\phi(\tilde{x}_{t+1})^T(x_t - x_{t+1}) \cdots \text{(By Bregman divergence)}$$

$$\leq \nabla\phi(x_t)^T(x_t - x_{t+1}) - \frac{\rho}{2}\|x_t - x_{t+1}\|^2 - \nabla\phi(\tilde{x}_{t+1})^T(x_t - x_{t+1})$$

$\cdots$ (ϕ is ρ-strongly convex)

(Cont.)

$$= \left(\nabla\phi(x_t) - \nabla\phi(\tilde{x}_{t+1})\right)^\top (x_t - x_{t+1}) - \frac{\rho}{2}\|x_t - x_{t+1}\|^2$$

$$= \eta_t \cdot \nabla f(x_t)^\top \cdot (x_t - x_{t+1}) - \frac{\rho}{2}\|x_t - x_{t+1}\|^2 \quad \cdots \text{(By mirror descent update)}$$

$$\leq \eta_t \cdot L_f \cdot \|x_t - x_{t+1}\| - \frac{\rho}{2}\|x_t - x_{t+1}\|^2 \quad \cdots \text{(By "Holder's inequality")}$$

$$\leq \frac{(\eta_t \cdot L_f)^2}{2\rho} \quad \cdots \text{(By completing the square)}$$

Proof of Main Theorem

Step1: By taking the telescoping sum of $\eta_\tau(f(x_\tau)-f^*)$, we have

$$\sum_{\tau=0}^{t}\eta_\tau\cdot(f(x_\tau)-f^*)\le D_\phi(x^*\|x_0)-D_\phi(x^*\|x_{t+1})+\frac{L_f^2\cdot\sum_{\tau=0}^{t}\eta_\tau^2}{2\cdot\rho}$$

$$\le \sup_{x\in C} D_\phi(x\|x_0)+\frac{L_f^2\cdot\sum_{\tau=0}^{t}\eta_\tau^2}{2\cdot\rho}$$

Step2: Moreover, we have

$$f_{best}^t-f^*\le\frac{\sum_{\tau=0}^{t}\eta_\tau(f(x_\tau)-f^*)}{\sum_{\tau=0}^{t}\eta_\tau}$$