## Definitions

- **Differentiable**: $f: \mathbb{R}^d \to \mathbb{R}$ is differentiable if
$$f(\mathbf{y}) = f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{r(\mathbf{x}-\mathbf{y})}{\|\mathbf{x}-\mathbf{y}\|},$$
where $\lim_{\mathbf{v}\to\mathbf{0}} \frac{\|r(\mathbf{v})\|}{\|\mathbf{v}\|} = 0$.
- **Spectral norm**: $\|A\|_2 = \sup_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$ (largest eigenvalue).
- **Positive semi-definite**: $\forall \mathbf{x} \in \mathbb{R}^d$: $\mathbf{x}^\top A \mathbf{x} \geq 0$.
- $B$-**Lipschitz**: $\|f(\mathbf{x}) - f(\mathbf{y})\| \leq B\|\mathbf{x}-\mathbf{y}\| \Leftrightarrow \|\nabla f(\mathbf{x})\| \leq B$.
- **Convex set**: $\forall \mathbf{x}, \mathbf{y} \in X, \lambda \in [0,1]$: $\lambda\mathbf{x} + (1-\lambda)\mathbf{y} \in X$.
- **Convexity**: $\forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$ and $\forall \lambda \in [0,1]$,
  [1] $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$.
  [2] $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle$.
  [3] $\langle \nabla f(\mathbf{x}) + \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \geq 0$.
  [4] $\nabla^2 f(\mathbf{x})$ is positive semi-definite.
- $L$-**smoothness**: $\forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$,
  [1] $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x}-\mathbf{y}\|$.
  [2] $g(\mathbf{x}) := \frac{L}{2}\|\mathbf{x}\|^2 - f(\mathbf{x})$ is convex.
  [3] $f(\mathbf{y}) \leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{L}{2}\|\mathbf{x}-\mathbf{y}\|^2$ (canonical).
  [4] $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \leq L\|\mathbf{x}-\mathbf{y}\|^2$.
  [5] $\|\nabla^2 f(\mathbf{x})\|_2 \leq L$.
  [6] Coordinate-wise: $f(\mathbf{x} + \lambda\mathbf{e}_i) \leq f(\mathbf{x}) + \lambda\nabla_i f(\mathbf{x}) + \frac{L_i}{2}\lambda^2, \forall \lambda \in \mathbb{R}$.
  Relations: $[5] \Leftrightarrow [1] \Rightarrow [2] \Leftrightarrow [3] \Leftrightarrow [4]$ (If convex, all $\Leftrightarrow$).
- $\mu$-**strong convexity**: $\forall \mathbf{x}, \mathbf{y} \in \mathrm{dom}(f)$,
  [1] $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{x}-\mathbf{y}\|^2$.
  [2] $g(\mathbf{x}) := f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex.
  [3] $\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x}-\mathbf{y}\rangle \geq \mu\|\mathbf{x}-\mathbf{y}\|^2$ (needs proof).
  [4] $\mu$-SC $\Rightarrow$ PL inequality: $\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 \geq \mu(f(\mathbf{x}) - f^\star)$.
- **Subgradient**: $\mathbf{g} \in \partial f(\mathbf{x}) \Leftrightarrow f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y}-\mathbf{x}\rangle, \forall \mathbf{y} \in \mathrm{dom}(f)$.
- **Conjugate function**: $f^\star(\mathbf{y}) := \sup_{\mathbf{x}\in\mathrm{dom}(f)} \langle \mathbf{x}, \mathbf{y}\rangle - f(\mathbf{x})$.

## Lemmas

- **Cosine theorem**: All equivalent formulations,
  [1] $\|\mathbf{x}-\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\langle \mathbf{x}, \mathbf{y}\rangle$.
  [2] $\langle \mathbf{x}, \mathbf{y}\rangle = \frac{1}{2}\big(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2\big)$.
  [3] $\langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{z}\rangle = \frac{1}{2}\big(\|\mathbf{x}-\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{z}\|^2 - \|\mathbf{y}-\mathbf{z}\|^2\big)$.
- **Cauchy-Schwarz**: $|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|\|\mathbf{y}\|$.
- **Hölder's inequality** (special case): $|\langle \mathbf{x}, \mathbf{y}\rangle| \leq \|\mathbf{x}\|_1\|\mathbf{y}\|_\infty$.
- **Jensen's inequality** ($\varphi$ convex, $a_i \geq 0$): $\varphi\Big(\frac{\sum_{i=1}^m a_i\mathbf{x}_i}{\sum_{i=1}^m a_i}\Big) \leq \frac{\sum_{i=1}^m a_i\varphi(\mathbf{x}_i)}{\sum_{i=1}^m a_i}$.
- **Fenchel's inequality**: $\langle \mathbf{x}, \mathbf{y}\rangle \leq f(\mathbf{x}) + f^\star(\mathbf{x})$ $\Rightarrow \langle \mathbf{x}, \mathbf{y}\rangle \leq \frac{1}{2}\big(\|\mathbf{x}\|^2 + \|\mathbf{y}\|_\star^2\big)$.
- **Young's inequality** ($a, b \geq 0, \frac{1}{p} + \frac{1}{q} = 1$): $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$ $\Rightarrow \|\mathbf{x}\|\|\mathbf{y}\| \leq \frac{1}{2}\big(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2\big)$.
- $\frac{1}{\sqrt{d}}\|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \|\mathbf{x}\|_1 \leq \sqrt{d}\|\mathbf{x}\|_2$.
- $\|A\mathbf{x}\| \leq \|A\|_2\|\mathbf{x}\|$.
- $\|A\|_2 \leq \|A\|_F$.
- **Mean-value theorem** ($h$ cont. on $[a,b]$, diff. on $(a,b)$):
$$h'(c) = \frac{h(b)-h(a)}{b-a}, \quad \exists c \in (a,b).$$
- **Fund. theorem of calculus** ($h$ diff. on $[a,b]$, $h'$ cont. on $[a,b]$):
$$h(b) - h(a) = \int_a^b h'(t)\mathrm{d}t.$$
- $\left\|\int_0^1 \nabla h(t)\mathrm{d}t\right\| \leq \int_0^1 \|\nabla h(t)\|\mathrm{d}t$.
- $\int_0^1 c\mathrm{d}t = c, \quad \int_0^1 t\mathrm{d}t = \frac{1}{2}$.
- **Subgradient calculus**:
  [1] $h(\mathbf{x}) = \alpha f(\mathbf{x}) + \beta g(\mathbf{x}) \Rightarrow \partial h(\mathbf{x}) = \alpha \cdot \partial f(\mathbf{x}) + \beta \cdot g(\mathbf{x})$.
  [2] $h(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b}) \Rightarrow \partial h(\mathbf{x}) = A^\top \partial f(A\mathbf{x}+\mathbf{b})$.
  [3] $h(\mathbf{x}) = \max f_i(\mathbf{x}) \Rightarrow \partial h(\mathbf{x}) = \mathrm{conv}(\{\partial f_i(\mathbf{x}) \mid f_i(\mathbf{x}) = h(\mathbf{x})\})$.

## Optimality lemmas (assume convexity)

The constrained and non-differentiable cases are useful when the update rule contains an argmin.

- $\mathbf{x}^\star$ is a local minimum.
- $\nabla f(\mathbf{x}^\star) = \mathbf{0}$.
- Constrained: $\nabla f(\mathbf{x}^\star)^\top(\mathbf{x} - \mathbf{x}^\star) \geq 0, \forall \mathbf{x} \in X$.
- Non-differentiable: $\mathbf{0} \in \partial f(\mathbf{x}^\star)$.

## Common tricks

- **Rearrange the update rule** for an equality—e.g., $\nabla f(\mathbf{x}_t) = \frac{\mathbf{x}_t - \mathbf{x}_{t+1}}{\gamma_t}$.
- Define $h(t) := f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))$, where $h'(t) = \nabla f(\mathbf{x}+t(\mathbf{y}-\mathbf{x}))^\top(\mathbf{y}-\mathbf{x})$ and use with fundamental theorem of calculus,
$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f(\mathbf{x} + t(\mathbf{y}-\mathbf{x}))^\top(\mathbf{y}-\mathbf{x})\mathrm{d}t.$$
Or, mean-value theorem,
$$\nabla f(\mathbf{x} + c(\mathbf{y}-\mathbf{x}))^\top(\mathbf{y}-\mathbf{x}) = f(\mathbf{y}) - f(\mathbf{x}), \quad \exists c \in (0,1).$$
- Projection is **non-expansive**: $\|\Pi_X(\mathbf{x}) - \Pi_X(\mathbf{y})\| \leq \|\mathbf{x}-\mathbf{y}\|$.
- $\min_{1\leq t\leq T} f(\mathbf{x}_t) - f^\star \leq \frac{\sum_{t=1}^T \gamma_t(f(\mathbf{x}_t) - f^\star)}{\sum_{t=1}^T \gamma_t}$.
- **Telescoping sum** inequality: $\sum_{t=1}^T \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq \|\mathbf{x}_1 - \mathbf{x}^\star\|^2$.
- $f^\star \leq f(\mathbf{x}), \forall \mathbf{x} \in X$ can sometimes be useful to bound $f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - f^\star$.
- $\max\{a, b\} \leq a + b$ if $a, b \geq 0$.

## Expectation and variance for SGD

- $\mathrm{Var}[X] := \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$
- $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ $\Rightarrow \mathbb{E}\|\nabla f(\mathbf{x}_t, \xi_t)\|^2 = \|\nabla F(\mathbf{x}_t)\|^2 + \mathbb{E}\|\nabla f(\mathbf{x}_t, \xi_t) - \nabla F(\mathbf{x}_t)\|^2 \leq \|\nabla F(\mathbf{x}_t)\|^2 + \sigma^2$.
- **Law of total expectation**: $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X \mid Y]]$.
- **Law of total var.**: $\mathrm{Var}[Y] = \mathbb{E}_X[\mathrm{Var}_Y[Y \mid X]] + \mathrm{Var}_Y[\mathbb{E}_X[Y \mid X]]$.
- $\mathrm{Var}[X - Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] - 2\cdot\mathrm{Cov}(X, Y)$.
- $\mathrm{Var}[\alpha X] = \alpha^2\mathrm{Var}[X]$, $\mathrm{Var}[X + \beta] = \mathrm{Var}[X]$.

## Risk minimization

## Non-linear programming

## Gradient descent

- **Update rule**: $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma\nabla f(\mathbf{x})$.
- **VA**: $\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f^\star) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$.

■ 1st-order convexity on $(\mathbf{x}^\star, \mathbf{x}_t) \Rightarrow \nabla f(\mathbf{x}_t) = \frac{\mathbf{x}_t - \mathbf{x}_{t+1}}{\gamma} \Rightarrow$ Cosine theorem $\Rightarrow \mathbf{x}_t - \mathbf{x}_{t+1} = \gamma\nabla f(\mathbf{x}_t) \Rightarrow$ Telescoping sum.

- **Sufficient decrease**: $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$.

■ Smoothness on $(\mathbf{x}_{t+1}, \mathbf{x}_t) \Rightarrow \mathbf{x}_{t+1} - \mathbf{x}_t = -\frac{1}{L}\nabla f(\mathbf{x}_t)$.

- **Convergence results**: ($\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$)
  - ($B$-Lipschitz, convex, $\gamma := \frac{R}{B\sqrt{T}}$) $\frac{1}{T}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f^\star) \leq \frac{RB}{\sqrt{T}}$.

    ■ Apply bounds to VA and find $\gamma$ by 1st-order optimality.

- ($L$-smooth, convex, $\gamma := \frac{1}{L}$) $f(\mathbf{x}_T) - f^\star \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$
  - ■ Sufficient decrease to bound gradients of VA with telescoping sum.
- ($L$-smooth, $\mu$-SC, $\gamma := \frac{1}{L}$) $f(\mathbf{x}_T) - f^\star \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$
  - ■ Use $\mu$-SC to strengthen VA bound for squared norm $\Rightarrow$ Upper bound "noise" with $f^\star \leq f(\mathbf{x}_{t+1})$ and SD $\Rightarrow$ Smoothness on $(\mathbf{x}^\star, \mathbf{x}_T)$.

## Projected gradient descent

- **Update rule** ($X \subset \mathbb{R}^d$ is closed and convex):
$$\mathbf{y}_{t+1} = \mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = \Pi_X(\mathbf{y}_{t+1}) := \operatorname*{argmin}_{\mathbf{x} \in \mathcal{X}} \|\mathbf{x} - \mathbf{y}_{t+1}\|^2.$$
- ($\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$): $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top(\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.
- ■ Constrained 1st-order optimality $\Rightarrow$ Rearrange.
- ($\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$): $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.
- ■ Cosine theorem on previous.
- If $\mathbf{x}_{t+1} = \mathbf{x}_t$, then $\mathbf{x}_t = \mathbf{x}^\star$.
- ■ $(\mathbf{x} - \mathbf{x}_t)^\top(\mathbf{y}_{t+1} - \mathbf{x}_t) = (\mathbf{x} - \mathbf{x}_{t+1})^\top(\mathbf{y}_{t+1} - \mathbf{x}_{t+1}) \leq 0, \forall \mathbf{x} \in X$ to show that constrained 1st-order optimality holds.
- **Projected SD**:
- ■
- ($L$-smooth, convex, $\gamma := \frac{1}{L}$):
- ■

## Coordinate descent

- **Coordinate-wise SD**:
- ■
- **Convergence results** ($\mu$-PL, $\mathcal{L}$-CS, $\bar{L} = \frac{1}{d}\sum_{i=1}^d L_i$):
  - ($i \sim \mathrm{Unif}([d])$)
  - ■
  - ($i \sim \mathrm{Cat}(L_1/\sum_{j=1}^d L_j, \ldots, L_d/\sum_{j=1}^d L_j)$)
  - ■
  - ($i \in \operatorname{argmax}_{j \in [d]} |\nabla_j f(\mathbf{x}_t)|$)

## Nonconvex functions

- ($L$-smooth):
- ■
- **Trajectory analysis**: Optimize $f(\mathbf{x}) := \frac{1}{2}\left(\prod_{k=1}^d x_k - 1\right)^2$.
- $\frac{\partial f(\mathbf{x})}{\partial x_i} = (\prod_k x_k - 1)\prod_{k \neq i} x_k$ ($\nabla f(\mathbf{x}) = \mathbf{0}$ if 2 dims are $0$ or all $1$).
- $\frac{\partial^2 f(\mathbf{x})}{\partial x_i^2} = \left(\prod_{k \neq i} x_k\right)^2$.
- $\frac{\partial^2 f(\mathbf{x})}{\partial x_i \, \partial x_j} = 2\prod_{k \neq i} x_k \prod_{k \neq j} x_k - \prod_{k \neq i,j} x_k$, if $i \neq j$.
- $c$-**balanced**: Let $\mathbf{x} > \mathbf{0}$, $c \geq 1$. $\mathbf{x}$ is $c$-balanced if $x_i \leq c \cdot x_j, \forall i, j \in [d]$.
- If $\mathbf{x}_t$ is $c$-balanced, $\gamma > 0$, then $\mathbf{x}_{t+1}$ is $c$-balanced and $\mathbf{x}_{t+1} \geq \mathbf{x}_t$.
- ■
- If $\mathbf{x}$ is $c$-balanced, then for any $I \subseteq [d]$, we have
$$\prod_{k \notin I} x_k \leq c^{|I|}\left(\prod_{k=1}^d x_k\right)^{1 - |I|/d} \leq c^{|I|}.$$
- ■
- Let $\mathbf{x}$ be $c$-balanced and $\prod_k x_k \leq 1$, then
$$\|\nabla^2 f(\mathbf{x})\|_2 \leq \|\nabla^2 f(\mathbf{x})\|_F \leq 3dc^2.$$
Thus, $f$ is smooth along the whole trajectory of GD.
- ■
- **Convergence** ($\gamma = \frac{1}{3dc^2}$, $\mathbf{x}_0 > \mathbf{0}$ and $c$-balanced, $\delta \leq \prod_k x_{0,k} < 1$) $f(\mathbf{x}_T) \leq \left(1 - \frac{\delta^2}{3c^4}\right)^T f(\mathbf{x}_0)$.
- ■
- $\delta$ decays polynomially in $d$, so we must start $\mathcal{O}(1/\sqrt{d})$ from $\mathbf{x}^\star = \mathbf{1}$.

## Frank-Wolfe

- $\mathrm{LMO}_X(\mathbf{g}) := \operatorname{argmin}_{\mathbf{z} \in X} \mathbf{g}^\top \mathbf{z}$.
- **Update rule**:
$$\mathbf{s}_t = \mathrm{LMO}_X(\nabla f(\mathbf{x}_t))$$
$$\mathbf{x}_{t+1} = (1 - \gamma_t)\mathbf{x}_t + \gamma_t \mathbf{s}_t.$$
- If $X = \mathrm{conv}(\mathcal{A})$, then $\mathrm{LMO}_X(\mathbf{g}) \in \mathcal{A}$.
- Advantages: (1) Iterates are always feasible if $X$ is convex, (2) No projections, (3) Iterates have simple sparse representations as convex combination of $\{\mathbf{x}_0, \mathbf{s}_0, \ldots, \mathbf{s}_t\}$.
- LMO of unit $\ell_1$-ball: $\mathrm{LMO}(\mathbf{g}) = -\mathrm{sgn}(g_i)\mathbf{e}_i, i \in \operatorname{argmax}_{j \in [d]} |g_i|$.
- **Optimality gap**: $g(\mathbf{x}) := \nabla f(\mathbf{x})^\top(\mathbf{x} - \mathbf{s}), \mathbf{s} = \mathrm{LMO}_X(\nabla f(\mathbf{x}))$.
- $g(\mathbf{x}) \geq f(\mathbf{x}) - f^\star$.
- ■
- **Descent lemma**: $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \gamma_t g(\mathbf{x}_t) + \gamma_t^2 \frac{L}{2}\|\mathbf{s}_t - \mathbf{x}_t\|^2$.
- **Convergence** ($L$-smooth, convex, $X$ is compact, $\gamma_t = \frac{2}{t+2}$): $f(\mathbf{x}_T) - f^\star \leq \frac{2L}{T+1}\mathrm{diam}(X)^2$.
- ■ Lemma$-f^\star$ $\Rightarrow$ Use $g(\mathbf{x}) \geq f(\mathbf{x}) - f^\star$ $\Rightarrow$ Rearrange and induction.
- **Affine equivalence**: $(f, X)$ and $(f', X')$ are affinely equivalent if $f'(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b})$ and $X' = \{A^{-1}(\mathbf{x} - \mathbf{b}) \mid \mathbf{x} \in X\}$. Then,
$$\nabla f'(\mathbf{x}') = A^\top \nabla f(\mathbf{x}), \quad \mathbf{x} = A^{-1}(\mathbf{x} - \mathbf{b})$$
$$\mathrm{LMO}_{X'}(\nabla f'(\mathbf{x}')) = A^{-1}(\mathbf{s} - \mathbf{b}), \quad \mathbf{s} = \mathrm{LMO}_X(\nabla f(\mathbf{x})).$$
- **Curvature constant**:
$$C_{(f,X)} := \sup_{\substack{\mathbf{x}, \mathbf{s} \in X, \gamma \in (0,1] \\ \mathbf{y} = (1-\gamma)\mathbf{x} + \gamma\mathbf{s}}} \frac{1}{\gamma^2}\left(f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x})\right).$$
- **Affine invariant convergence**: $f(\mathbf{x}_T) - f^\star \leq \frac{4C_{(f,X)}}{T+1}$.
- ■ Descent lemma w.r.t. $C_{(f,X)}$ by setting $\mathbf{x} = \mathbf{x}_t, \mathbf{s} = \mathrm{LMO}_X(\nabla f(\mathbf{x}_t))$ in the supremum.
- **Convergence of** $g(\mathbf{x}_t)$: $\min_{1 \leq t \leq T} g(\mathbf{x}_t) \leq \frac{27/2 \cdot C_{(f,X)}}{T+1}$.

## Newton's method

- **Update rule**: $\mathbf{x}_{t+1} = \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t)$.
- **Interp. 1**: Adaptive gradient descent.
- **Interp. 2**: Minimizes second-order Taylor approximation around $\mathbf{x}_t$:
$$\mathbf{x}_{t+1} \in \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x} - \mathbf{x}_t) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t).$$
- **Convergence** ($\|\nabla^2 f(\mathbf{x})^{-1}\| \leq \frac{1}{\mu}$, $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq B\|\mathbf{x} - \mathbf{y}\|$): $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq \frac{B}{2\mu}\|\mathbf{x}_t - \mathbf{x}^\star\|^2$.
- ■ $\mathbf{x}_{t+1} - \mathbf{x}^\star \leq \mathbf{x}_t - \mathbf{x}^\star + H(\mathbf{x}_t)^{-1}(\nabla f(\mathbf{x}^\star) - \nabla f(\mathbf{x}_t)) \Rightarrow h(t) := \nabla f(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x}))$ with fundamental theorem of calculus $\Rightarrow$ Take norm of both sides and simplify using $\|A\mathbf{x}\| = \|A\|_2\|\mathbf{x}\|$ and assumptions.
- Ensure bounded inverse Hessians by requiring strong convexity over $X$.
- If $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq \frac{\mu}{B}$, then $\|\mathbf{x}_T - \mathbf{x}^\star\| \leq \frac{\mu}{B}\left(\frac{1}{2}\right)^{2^T - 1}$.

## Quasi-Newton methods

## Subgradient method

## Mirror descent

## Stochastic optimization

## Variance reduction

## Min-max optimization

## Variational inequalities