Gradient Descent #2

SGD, On Data
\[(X, Y) = \{ (x_i, y_i), \forall i \in [n] \}\]

\[
\{ \mathbf{m}_i | \mathbf{a} \in \mathbb{R}^m \}
\]

\[
\mathbf{x}_i \in \mathbb{R}^d
\]

\[
\mathbf{y}_i \in \mathbb{R}^n
\]

\[
\text{Argmin}_{\mathbf{a} \in \mathbb{R}^m} \frac{1}{2} \left( \mathbf{x}_i - \mathbf{y}_i \right)^T \mathbf{M} \left( \mathbf{x}_i - \mathbf{y}_i \right) + \frac{1}{2} \mathbf{a}^T \mathbf{x}_i \mathbf{a}
\]

\[
\text{Argmin}_{\mathbf{a} \in \mathbb{R}^m} \mathbf{f}(\mathbf{a})
\]
\[(x_i, y_i) = \{ (x_i, y_i), \ldots, (x_m, y_m) \}\]

\[x_i \in \mathbb{R} \quad (d=1) \quad y_i \in \mathbb{R}\]

Poly. reg. \[p = 2\]

\[x_i \rightarrow (1, x_i, x_i^2)\]

\[
\alpha = (d_0, d_1, d_2) \in \mathbb{R}^3
\]

\[\mathcal{M}_\alpha(x_i) = d_0 + d_1 x_i + d_2 x_i^2 = \langle d_0, 1, x_i, x_i^2 \rangle \]

\[\text{SSE}(x_i, y_i, \mathcal{M}_\alpha) = \sum_{i=1}^{n} (\mathcal{M}_\alpha(x_i) - y_i)^2 \]

\[\text{feasible} \quad = \sum_{i=1}^{n} (y_i - \mathcal{M}_\alpha(x_i))^2\]
\[ f(\alpha) = \frac{1}{N} \sum_{i=1}^{N} f_i(\alpha) \]

\[ f_i(\alpha) = (M_{\alpha}(x_i) - y_i) \]

\[ \alpha^* = \arg \min_{\alpha \in \mathbb{R}^3} f(\alpha) \]

\[
\begin{align*}
\alpha^{(0)} &= \alpha^* \\
\text{repeat} & \\
\alpha^{(k+1)} &= \alpha^{(k)} - \gamma_k \nabla f(\alpha^{(k)}) \\
\text{until } & \\
\end{align*}
\]
\[ f(\alpha) = \sum_{i=1}^{N} f_i(\alpha) \]
\[ f_i(\alpha) = (M \alpha (x_i - y_i))^2 \]
\[ \alpha = (\alpha_0, \alpha_1, \alpha_2) \]
\[ N = 1 \]
\[ (x_i, y_i) \]
\[ f(\alpha) = (M \alpha (x_i - y_i))^2 \]
\[ \nabla f(\alpha) = \left( \frac{\partial}{\partial \alpha}, \frac{\partial}{\partial \alpha_1}, \frac{\partial}{\partial \alpha_2} f(\alpha) \right) \]
\( f_i(x_1) = (M_a(x_1) - y_i)^2 \)

\[ \frac{\partial}{\partial x_2} f = 2(M_a(x_1) - y_i) x_i \]

\[ \frac{\partial}{\partial a} \left( (M_a(x_1) - y_i)^2 \right) = 2(M_a(x_1) - y_i) \frac{\partial (M_a(x_1) - y_i)}{\partial a} \]

\[ = 2 (M_a(x_1) - y_i) \frac{\partial (a + \alpha x_1 + \delta x_1^2 - y_i)}{\partial a} \]

\[ = 2 (M_a(x_1) - y_i) \cdot 1 \]

\[ \frac{\partial}{\partial x_1} \left( (M_a(x_1) - y_i)^2 \right) = 2(M_a(x_1) - y_i) \frac{\partial (a + \alpha x_1 + \delta x_1^2 - y_i)}{\partial x_1} \]

\[ = 2 (M_a(x_1) - y_i) x_i \]
\[ N = 1 \]
\[ \nabla f(\alpha) = 2 (M_{\alpha}(x_i) - y_i)(1, x_i, x_i^2) \]

\[ \alpha_{\text{new}} = \alpha_{\text{old}} - \gamma 2 (M_{\alpha_{\text{old}}}(x_i) - y_i)(1, x_i, x_i^2) \]

\[ N > 1 \quad \rightarrow \quad \text{decomposable} \]

\[ f(\alpha) = \sum_{i=1}^{n} f_i(\alpha) \]

\[ f_i(\alpha) = (M_{\alpha}(X_i) - y_i)^2 \]
\[ \nabla f(\alpha) = \sum_{i=1}^{N} f_i(\alpha) \]
\[ = \sum_{i=1}^{N} \nabla f_i(\alpha) \]

\[ \nabla f_i(\alpha) = 2 (\nabla \alpha_i (x_i) - y_i) (1, x_i, x_i^2) \]

\[ \nabla f(\alpha) = \sum_{i=1}^{N} 2 (\nabla \alpha_i (x_i) - y_i) (1, x_i, x_i^2) \]
\[ p = 1 \]

\[ f(a) = \sum_{i=1}^{n} f_i(a) = \sum_{i=1}^{n} (M_a(x_i) - y_i)^2 \]

\[ = \sum_{i=1}^{n} (\alpha_0 + \alpha_1 x_i - y_i)^2 \]

If each \( f_i \) is convex, then \( \sum f_i \) is convex.

\[ f_i(a) = 0 \iff \alpha_0 + \alpha_1 x_i - y_i = 0 \]

\[ \iff \alpha = y_i - \alpha x_i \]
\[ f_1(\alpha_1) = (\mu \alpha (x_1 - y_1))^2 \]

\[ a_0 = y_2 - a_1 x_2 \]

\[ a = y_1 - \alpha x_1 \]
\[ f(x) = \sum_{i} f_i(x) \text{ is strongly convex} \]

\[
\text{if } N \geq 2 \text{ (if at least two points are not non-general)}
\]

If \( N \geq \#\text{parameters} \Rightarrow f \text{ is strongly convex.} \]
\[ \nabla f(x) = \sum_{i=1}^{N} \nabla f_i(x) \]

---

**Incremental gradient descent**

init. \( a^{(0)} = a^{\text{Start}} \)

repeat

\[ a^{(k+1)} = a^{(k)} - \gamma \nabla f_i(a^{(k)}) \]

\( i = (i + 1 \mod N) \)

until \( \| \nabla f_i(a) \| < \tau \)

\( \Rightarrow \) average of \( \| \nabla f_i(a) \| < \tau \) for some iterations.
Stochastic gradient descent

initialize $\mathbf{d}^{(0)} = \mathbf{a}^{\text{start}}$

repeat

randomly select $i \in \{1, 2, \ldots, n\}$

$$
\mathbf{d}^{(\text{alt})} = \mathbf{a}^{(k)} + \eta \nabla f_i (\mathbf{d}^{(k)})
$$

until $(\|\nabla f(\mathbf{d})\|_{\infty} < \epsilon)$
\[ \mathcal{D} f (x) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{D} f_i (x) \]

\[ \mathbb{E} ( \mathcal{D} f_i (x) ) = \mathcal{D} f (x) \]

\[ i \sim \{ 1, \ldots, N \} \]