Recap

$$\min_x \; f(x) \quad \text{such that} \quad x \in C$$

$\longrightarrow$ GD + Projection $\quad x^{(k)} = P_C\left(x^{(k-1)} - t_k g^{(k-1)}\right)$

Let $I_C(x) = \begin{cases} \infty & x \notin C \\ 0 & x \in C \end{cases}$ then

$$\min_x \; f(x) + I_C(x) \quad \text{is equivalent}$$

Proximal Gradient Descent

$$f(x) = g(x) + h(x)$$

$g(x)$: convex & differentiable

$h(x)$: convex but non-smooth

e.g. $\min_x f(x)$ such that $\|x\|_1 < \varepsilon$

$(\|x\|_1 - \varepsilon < 0)$

Lagrange Mult. $\longrightarrow \min f(x) + \lambda(\|x\|_1 - \varepsilon)$

$$\equiv \min f(x) + \lambda \|x\|_1$$

$g(x)$ $\quad$ $h(x)$

Want to $\min_z \; g(z) + h(z)$

$$\therefore \quad g(z) \lesssim g(x) + \nabla g(x)^T (z-x) + \frac{1}{2t} \|z-x\|_2^2$$

$(\lesssim)$ $\quad \left(\nabla^2 g(x) \preceq LI \;\; \& \;\; t \leq \frac{1}{L} \right)$

$$\Rightarrow \quad \min_z \; g(z) + h(z)$$

$$\cong \quad \min_z \; g(x) + \nabla g(x)^T (z-x) + \frac{1}{2t} \|z-x\|_2^2 + h(z)$$

$$\cong \quad \min_z \; \frac{1}{2t} \|z - (x - t\nabla g(x))\|_2^2 + h(z)$$

close to gradient of g ; minimize h

Definition Proximal Operator

$$\text{prox}_{h,t}(x) = \underset{z}{\text{argmin}} \; \|x-z\|_2^2 + h(z)$$

Proximal Update : $x^{(k)} = \text{prox}_{h,t_k}\left(x^{(k-1)} - t_k \nabla g(x^{(k-1)}) \right)$

Let $G_t(x) = \dfrac{x - \text{prox}_{h,t}(x - t\nabla g(x))}{t}$

then $x^{(k)} = x^{(k-1)} - t_k G_{t_k}(x^{(k-1)})$

is proximal Update

Going back to $\min_x f(x)$ s.t. $x \in C$

$\equiv \min_x f(x) + I_C(x)$

$\text{prox}_{I_C,t} = \underset{z}{\text{argmin}} \; \|x - z\|_2^2 + I_C(z)$

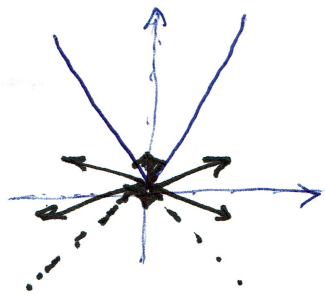
↳ which is projection to C

Example: Iterative Soft Thresholding Algorithm

LASSO: $\min_\beta \underbrace{\frac{1}{2}\|y - X\beta\|_2^2}_{g(\beta)} + \underbrace{\lambda\|\beta\|_1}_{h(\beta)}$

$$\text{prox}_{h,t}(\beta) = \underset{z}{\text{argmin}} \; \frac{1}{2t}\|x-z\|_2^2 + \lambda\|z\|_1$$

$$\partial\|z\|_1 = \begin{cases} \text{sgn}(z) & z \neq 0 \\ \alpha \in [-1,1] & z = 0 \end{cases}$$



Rewrite : $\underset{z}{\text{argmin}} \; \frac{1}{2}\|x-z\|_2^2 + \lambda t\|z\|_1$

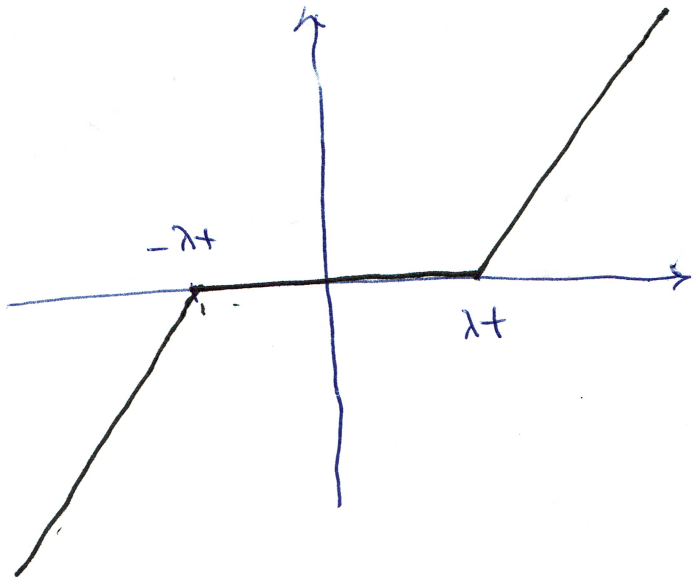
$$0 \equiv \frac{\partial}{\partial z} \Rightarrow 0 \equiv z - x + \lambda t \, \partial\|z\|_1$$

$$\begin{cases} \text{if } |x_i| \leq \lambda t & \text{Case I} \\ \text{if } |x_i| > \lambda t & \text{Case II} \end{cases}$$

Case I $\Rightarrow$ $z_i = 0$; $\partial\|z\|_{1,i} = \frac{x}{\lambda t}$

Case II $\Rightarrow$ $z_i = x_i - \lambda t$; $\partial\|z\|_{1,i} = \text{Sgn}(x_i)$

$$S_{\lambda t}(\beta) \equiv \text{prox}_{h,t}(\beta) = \begin{cases} \beta_i - \lambda t & \beta_i > \lambda t \\ 0 & -\lambda t \leq \beta_i \leq \lambda t \\ \beta_i + \lambda t & \beta_i < -\lambda t \end{cases}$$



$\text{prox}_{h,t}(x)$

$$\Rightarrow \quad \beta^{(k)} = S_{\lambda t}\left(\beta^{(k-1)} - t\, X^T(y - X\beta^{(k-1)})\right)$$

Convergence: $f(x^{(k)}) - f^* \leq \dfrac{\|x^{(0)} - x^*\|_2^2}{2tk}$

Accellerated Proximal Gradient Descent

$$\min_x \; g(x) + h(x)$$

$$\begin{cases} v = x^{(k-1)} + \dfrac{k-2}{k+1}\left(x^{(k-1)} - x^{(k-2)}\right) \\ x^{(k)} = \text{prox}_{h,t_k}\left(v - t_k \nabla g(v)\right) \end{cases}$$

Convergence:

$$f(x^{(k)}) - f^* \leq \frac{2\|x^{(0)} - x^*\|_2^2}{t(k+1)^2}$$

↳ acceleration improvement

---

practical Notes

Arcelleration vs. warm Start

$$\lambda_1 > \lambda_2 > \cdots \geq \lambda_r$$

Let $x^{(0)} = x^*(\lambda_{j-1})$

Often warm start does just as well notice $\|x^{(0)} - x^*\|_2^2$ is always a factor in convergence with limited # of steps

Stochastic Gradient Descent

Let's go back to $\min_x f(x)$ problem.

Often, $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$

↳ observations

---

example: Regression $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

$$f(\beta) = \frac{1}{n} \sum_{i=1}^{n} \| y_i - X_i \beta \|_2^2$$

↳ mean squared error

Gradient Descent: $x^{(k)} = x^{(k-1)} - t_k \frac{1}{n} \sum_{i=1}^{n} \nabla f_i$

Stochastic " " pick $i_k \in \{1, \ldots, n\}$ at random

$$x^{(k)} = x^{(k-1)} - t_k \nabla f_{i_k}(x^{(k-1)})$$

↳ "only a single term"

We hope that

$$\mathbb{E}\left[\nabla f_{i_k}(x^{(k-1)})\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^{(k-1)})$$

Mini batch : pick $1 \leq b \leq n$

and $I_k \subseteq \{1, \dots, n\}$ s.t. $|I_k| = b$

(I_k: random)

$$x^{(k)} = x^{(k-1)} - t_k \frac{1}{b} \sum_{i \in I_k} \nabla f_i(x^{(k-1)})$$

there is higher chance that

$$\mathbb{E}\left[\frac{1}{b}\sum_{i \in I_k} \nabla f_i(x^{(k-1)})\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^{(k-1)})$$

	GD	SGD
convex	$O(\frac{1}{\sqrt{k}})$	$O(\frac{1}{\sqrt{k}})$
Lipschitz	$O(\frac{1}{k})$	$O(\frac{1}{\sqrt{k}})$
Strongly Convex	$O(\gamma^k)$	$O(1/k)$

Theorem f is convex and M-Liepshitz

$$\left(f(y) \le f(x) + \nabla f(x)(y-x) + \frac{M}{2}\|y-x\|_2^2\right)$$

OR: $\nabla^2 f \le MI$

Let $t_k \le \dfrac{1}{M + \sigma\frac{\sqrt{k}}{\|x^{(0)}-x^*\|_2}} = \dfrac{\|x^{(0)}-x^*\|_2}{M\|x^{(0)}-x^*\|_2 + \sigma\sqrt{k}}$

with $\sigma^2 = \max_x \|\hat{\nabla} f(x) - \nabla f(x)\|_2^2$

how good our estimate of gradient is

notice if $\sigma = 0$ then $t_k \le \frac{1}{M}$ (exactly GD)

Then $\mathbb{E}\left[\frac{1}{k}\sum f(x^{(k)})\right] - f(x^*) \le$

$$\frac{M\|x^{(0)}-x^*\|_2^2}{k} + \frac{\sigma\|x^{(0)}-x^*\|_2}{\sqrt{k}}$$

notice if $\sigma = 0$ then this is exactly GD error bound

proof

$$f(x^{(k+1)}) \le f(x^{(k)}) + \nabla f(x^{(k)})(x^{(k+1)} - x^{(k)}) + \frac{M}{2}\|x^{(k+1)} - x^{(k)}\|_2^2$$

by convexity:

$$f(x^*) \ge f(x^{(k)}) + \nabla f(x^{(k)})(x^* - x^{(k)})$$

Hence:

$$f(x^{(k+1)}) - f(x^*) \le \nabla f(x^{(k)})(x^{(k)} - x^*) + \nabla f(x^{(k)})(x^{(k+1)} - x^{(k)}) + \frac{M}{2}\|x^{(k+1)} - x^{(k)}\|_2^2$$

By SGD step, $x^{(k+1)} = x^{(k)} - t_k \hat{\nabla} f(x^{(k)})$; hence

$$f(x^{(k+1)}) - f(x^*) \le \nabla f(x^{(k)})(x^{(k)} - x^*) + \nabla f(x^{(k)})(-t_k \hat{\nabla} f(x^{(k)})) + \frac{M t_k^2}{2}\|\hat{\nabla} f(x^{(k)})\|_2^2$$

$$= (\nabla f(x^{(k)}) - \hat{\nabla} f(x^{(k)}))(x^{(k)} - x^*) + \hat{\nabla} f(x^{(k)})(x^{(k)} - x^*) + (\nabla f(x^{(k)}) - \hat{\nabla} f(x^{(k)}))(-t_k \hat{\nabla} f(x^{(k)})) + \hat{\nabla} f(x^{(k)})(-t_k \hat{\nabla} f(x^{(k)})) + \frac{M t_k^2}{2}\|\hat{\nabla} f(x^{(k)})\|_2^2$$

$$(\nabla - \hat{\nabla})(x^{(k)} - x^*) + (\nabla - \hat{\nabla})(-t_k \hat{\nabla})$$

$$= (\nabla - \hat{\nabla})(x^{(k)} - x^* - t_k \nabla) + t_k (\nabla - \hat{\nabla})(\nabla - \hat{\nabla})$$

$(\nabla - \hat{\nabla})(x^{(k)} - x^* - t_k \nabla) = 0$

$(\nabla - \hat{\nabla})(\nabla - \hat{\nabla}) = \sigma^2$

$$\Rightarrow \mathbb{E}\left[(\nabla - \hat{\nabla})(x^{(k)} - x^*) + (\nabla - \hat{\nabla})(-t_k \hat{\nabla})\right] = t_k \sigma^2$$

Notice:

$$\frac{1}{2t_k} \| x^{(k+1)} - x^* \| = \frac{1}{2t_k} \| x^{(k)} - x^* - t_k \hat{\nabla} \|_2^2$$

$$= \frac{1}{2t_k} \| x^{(k)} - x^* \|_2^2 + \frac{t_k}{2} \| \hat{\nabla} \|_2^2 - \hat{\nabla}(x^{(k)} - x^*) \quad \text{(I)}$$

NOW: the rest of the terms

$$\hat{\nabla}(x^{(k)} - x^*) - t_k \| \hat{\nabla} \|_2^2 + \frac{M t_k^2}{2} \| \hat{\nabla} \|_2^2$$

$$\leq \hat{\nabla}(x^{(k)} - x^*) - \frac{t_k}{2} \| \hat{\nabla} \|_2^2 \quad \text{assumption on } t_k$$

$$\overset{\text{(I)}}{=} \frac{1}{2t_k} \left[\| x^{(k)} - x^* \|_2^2 - \| x^{(k+1)} - x^* \|_2^2 \right]$$

Finally

$$f(x^{(k+1)}) - f(x^*) \leq \frac{1}{2t_k}\left[\|x^{(k)} - x^*\|_2^2 - \|x^{(k+1)} - x^*\|_2^2\right] + t_k\sigma^2$$

$$\Rightarrow \mathbb{E}\left[\frac{1}{k}\sum f(x^{(k)})\right] - f(x^*) \leq \frac{\|x^{(0)} - x^*\|_2^2}{2t_k k} + t_k\sigma^2$$

plugging $t_k \leq \frac{1}{M + \frac{\sigma\sqrt{k}}{\|x^{(0)} - x^*\|_2}}$ we get the bound

---

Variance Reduction

we want "σ" to be small !

stochastic Average Gradient (SAG)

→ Initialize $g_i^{(0)} = \nabla f_i(x^{(0)})$

→ Randomly pick $i_k \in \{1, \dots, n\}$ and $\begin{cases} g_{i_k}^{(k)} = \nabla f_{i_k}(x^{(k-1)}) \\ g_{\neq i_k}^{(k)} = g_{\neq i_k}^{(k-1)} \end{cases}$

→ $\hat{\nabla} = \frac{1}{n}\sum_{i=1}^{n} g_i^{(k)}$

SAG Convergence: $t_k \leq \frac{1}{16M}$; $g_i^{(0)} = \nabla f_i(x^{(0)})$ - (13) $\nabla f(x^{(0)})$

$$\mathbb{E}\left[\frac{1}{k}\sum f(x^{(k)})\right] - f(x^*) \leq \frac{48n}{k}\left(f(x^{(0)}) - f(x^*)\right)$$

$$+ \frac{128M}{k}\|x^{(0)} - x^*\|_2^2$$

Does NOT depend on "σ"

SAG, strong convexity convergence:

$$\mathbb{E}\left[\frac{1}{k}\sum f(x^{(k)})\right] - f(x^*) \leq \gamma(m, M, n)^k \left(\frac{3}{2} f(x^{(0)}) - f(x^*)\right)$$

$$+ \frac{4M}{n}\|x^{(0)} - x^*\|_2^2$$

exponential ← BUT ↓ indep. of "k"

SAGA

$$\hat{\nabla} = g_{i_k}^{(k)} - g_{i_k}^{(k-1)} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} g_i^{(k-1)}}_{\text{SAG}}$$

AdaGrad (Adam)

$$\hat{\nabla}_j = \frac{g_j^{(k)}}{\sqrt{\sum_{q=1}^{k} (g_j^{(q)})^2}}$$

We hope that

$$\mathbb{E}\left[\nabla f_{i_k}(x^{(k-1)})\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^{(k-1)})$$

Mini batch : pick $1 \leq b \leq n$

and $I_k \subseteq \{1, \dots, n\}$ s.t. $|I_k| = b$

(I_k: random)

$$x^{(k)} = x^{(k-1)} - t_k \frac{1}{b} \sum_{i \in I_k} \nabla f_i(x^{(k-1)})$$

there is higher chance that

$$\mathbb{E}\left[\frac{1}{b}\sum_{i \in I_k} \nabla f_i(x^{(k-1)})\right] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x^{(k-1)})$$

	GD	SGD
convex	$O(\frac{1}{\sqrt{k}})$	$O(\frac{1}{\sqrt{k}})$
Lipschitz	$O(\frac{1}{k})$	$O(\frac{1}{\sqrt{k}})$
Strongly Convex	$O(\gamma^k)$	$O(1/k)$