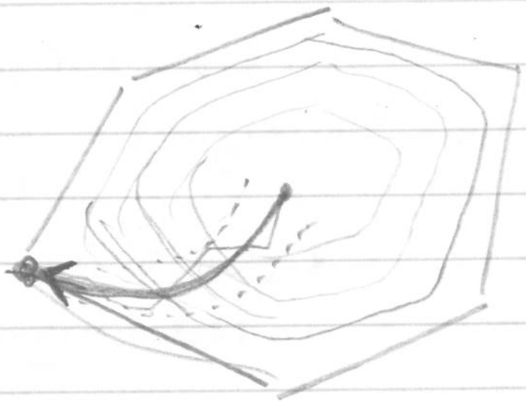

$$g(x) + \nabla g(x)^T (z-x) + \frac{1}{2t}\|z-x\|_2^2$$

Gradient

$$g(x) + \nabla g(x)^T(z-x) + \frac{1}{2}(z-x)^T \nabla^2 g(x)(z-x)$$

Newton

H-norm: $\|x\|_H^2 = x^T H x$

$$x^{(k)} = x^{(k-1)} + t_k(y^{(k)} - x^{(k-1)})$$

E.g. $\min_\beta \; l(\beta) + \lambda\|\beta\|_1$

logistic loss.

not a good tradeoff.

Prox gradient directly: require $O(1/\varepsilon)$ iterations

Prox Newton. with prox grad to evaluate scaled prox. ?
$O(1/\varepsilon)$ iterations each time! + outer iterations

Why was this a bad idea?

Because prox gradient applies equally well to quadratic & logistic loss ($O(1/\varepsilon)$ in both cases)

<u>So</u>: use a method in inner loop that's specific to <u>quadratic loss</u>. In practice: coordinate descent.

much faster than $O(1/\varepsilon)$ for quadratic loss

$$mI \preceq \nabla^2 g \preceq LI$$

$$\min_{x} \quad g(x) + h(x)$$

Tseng and Yun: outer loop: over blocks $b = 1, .. B$

inner loop: apply quadratic approx to $g(x)$ over x_b

Proximal Newton: outer loop: apply quad approx to $g(x)$ over x

inner loop: over blocks $b = 1 .. B$