Newton Method

Second Order Approximation:

$$f(x_k + \alpha p_k) \cong f(x_k) + \underbrace{\alpha \nabla f(x_k)^T p_k + \frac{\alpha^2}{2} p_k^T \nabla^2 f(x_k) p_k}_{\overset{\text{def}}{=} \phi(p_k)}$$

To find the "best" search direction, we want to find a p_k s.t. $f(x_k + \alpha p_k)$ is minimized.

For simplicity, let's assume $\alpha = 1$. Then

$$\frac{\partial}{\partial p_k} \phi(p_k) = \nabla f(x_k) + \nabla^2 f(x_k) p_k = 0$$

$$\Rightarrow \quad \boxed{p_k = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)} \quad \text{———} (1)$$

Assumption: $\nabla^2 f(x_k) \succ 0$.

When will (1) be valid? (1) is valid if $\phi(p_k)$ is a good approximation of $f(x_k + \alpha p_k)$, which happens only when αp_k is sufficiently close to x_k.

if $p_k = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$, then one can show that

$$\nabla f(x_k)^T p_k = -\nabla f(x_k)^T \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$$

$$< 0 \qquad , \qquad = 0 \text{ when } \nabla f(x_k) = 0.$$

Step size α: We can actually fix $\alpha = 1$! or using line search

gradient descent: $\min_p \nabla f(x_k)^T p$ s.t. $\|p\|_2 = 1$ makes sure minimization has solution but loose length

Newton : $\min_p \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p$ $\nabla^2 f(x_k)$ is defining the length

Compare Newton & descent

$$p_k = \underbrace{-\left[\nabla^2 f(x_k)\right]^{-1}}_{\text{vector step size}} \nabla f(x_k) \qquad p_k = -\left[\underbrace{\alpha_k}_{\text{scalar step size}} I\right] \nabla f(x_k)$$

Convergence of Newton

Assumption: (1) f is twice differentiable so that $\nabla^2 f$ exists.

(2) $\nabla^2 f(x)$ is Lipschitz continuous in a neighborhood of x^*:

$$\| \nabla^2 f(x) - \nabla^2 f(y) \| \leq L \| x - y \| \quad \forall x, y \in \text{this neighbor}$$

Theorem (3.7 Nocedal & Wright)

(1) if x_0 is sufficiently close to x^*, then $\{x_k\}$ converges to x^*

(2) rate of convergence of $\{x_k\}$ is quadratic:

$$\| x_{k+1} - x^* \| \leq C \| x_k - x^* \|^2, \text{ some constant } C$$

(3) rate of convergence of $\{\nabla f(x_k)\}$ is quadratic:

$$\| \nabla f(x_{k+1}) - 0 \| \leq C' \| \nabla f(x_k) - 0 \|^2, \text{ some constant } C'.$$

Proof: First, note that

$$\begin{aligned}
x_{k+1} - x^* &= x_k + p_k - x^* \\
&= x_k - \left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k) - x^* \\
&= \left[\nabla^2 f(x_k)\right]^{-1} \left[\nabla^2 f(x_k)(x_k - x^*) - \nabla f(x_k) \right] \\
&= \left[\nabla^2 f(x_k)\right]^{-1} \left[\nabla^2 f(x_k)(x_k - x^*) - \left(\nabla f(x_k) - \underbrace{\nabla f(x^*)}_{=0}\right) \right]
\end{aligned}$$

Mean-Value theorem:

if f is continuous on $[a,b]$, then $\exists c$ s.t. $a \le c \le b$,

$$f(c)(b-a) = \int_a^b f(x)\,dx$$

Multi-variate: (see Wikipedia)

$$f(x+p) - f(x) = \int_0^1 \nabla f(x+tp)^T p \, dt$$

Fund. theorem of Calculus:

$$f(x+p) - f(x) = \int_x^{x+p} f'(u)\,du = \int_0^1 f'(x+tp)\,dt\; p$$

Therefore,

$$\nabla f(x_k) - \nabla f(x^*) = \int_0^1 \nabla^2 f(x_k + t(x^* - x_k))\,dt\,(x_k - x^*)$$

$$= \int_0^1 \nabla^2 f\left(x_k + t(x^* - x_k)\right) \cdot (x_k - x^*)\,dt.$$

So,

$$\| \nabla^2 f(x_k)(x_k - x^*) - (\nabla f(x_k) - \nabla f(x^*)) \|$$

$$= \| \nabla^2 f(x_k)(x_k - x^*) - \int_0^1 \nabla^2 f(x_k + t(x^* - x_k))(x_k - x^*)\,dt \|$$

$$= \| \int_0^1 \left[\nabla^2 f(x_k) - \nabla^2 f(x_k + t(x^* - x_k)) \right](x_k - x^*)\,dt \|$$

$$\le \int_0^1 \| \nabla^2 f(x_k) - \nabla^2 f(x_k + t(x^* - x_k)) \| \, \| x_k - x^* \|\,dt$$

$$\le \int_0^1 L \| x_k - (x_k + t(x^* - x_k)) \| \, \| x_k - x^* \|\,dt$$

$$= \cancel{L} \| x_k - x^* \|^2 \int_0^1 L\,t\,dt = \frac{L}{2} \| x_k - x^* \|^2$$

Banach Perturbation Lemma:

A : non-singular, $\|A^{-1}\| \le \mu$. If $\|A - \bar{A}\| \le \epsilon$, where $\epsilon\mu < 1$, then $\bar{A}$ is non-singular and

$$\|\bar{A}^{-1}\| \le \frac{\mu}{1-\mu\epsilon}.$$

$\epsilon = \frac{1}{2\mu}$

<u>Corollary</u>: $\nabla^2 f(x^*) \succ 0$. Let $\mu = \|\nabla^2 f(x^*)^{-1}\|$. Then there is a $\mathcal{B}(x^*, \delta)$ s.t. $\forall x \in \mathcal{B}(x^*, \delta)$,

$$\|\nabla^2 f(x)^{-1}\| < 2\mu = 2\|\nabla^2 f(x^*)^{-1}\|.$$

Then,

$$\|x_{k+1} - x^*\| \le \underbrace{L\|\nabla^2 f(x^*)^{-1}\|}_{C} \|x_k - x^*\|^2.$$

So $\{x_k\}$ converges quadratically to x^*. this shows (a) and (b).

For (c),

$$\|\nabla f(x_{k+1}) - 0\| = \|\nabla f(x_{k+1})\|$$

$$= \|\nabla f(x_{k+1}) - \boxed{(\nabla f(x_k) + \nabla^2 f(x_k) p_k)}\|$$

$=0$ by Newton method definition

$$\cancel{= \|\int_0^1 \nabla^2 f(x_{k+1})}$$

$$= \left\|\int_0^1 \nabla^2 f(x_k + tp_k)(x_{k+1} - x_k)\, dt - \nabla^2 f(x_k) p_k\right\|$$

$$\cancel{\le \int_0^1 \|\nabla^2 f(x_k + tp_k)(x_{k+1} - x_k)\, dt\|}$$

$$= \left\|\int_0^1 [\nabla^2 f(x_k + tp_k) - \nabla^2 f(x_k)] p_k\, dt\right\|$$

$$\le \int_0^1 \|\nabla^2 f(x_k + tp_k) - \nabla^2 f(x_k)\| \|p_k\|\, dt$$

$$\le \frac{1}{2} L \|p_k\|^2$$

$$= \frac{1}{2} L \| [\nabla^2 f(x_k)]^{-1} \nabla f(x_k) \|^2$$
$$\leq \frac{1}{2} L \| \nabla^2 f(x_k)^{-1} \|^2 \| \nabla f(x_k) \|^2$$
$$\leq \frac{1}{2} L \; 4 \| \nabla^2 f(x^*)^{-1} \|^2 \| \nabla f(x_k) \|^2$$
$$= 2L \| \nabla^2 f(x^*)^{-1} \|^2 \| \nabla f(x_k) \|^2.$$

So we conclude that

$$\| \nabla f(x_{k+1}) \| \leq \underbrace{2L \| \nabla^2 f(x^*)^{-1} \|^2}_{C'} \| \nabla f(x_k) \|^2.$$

Quasi-Newton Method

Recall Mean-Value Theorem:

$$f(x+p) - f(x) = \int_0^1 \nabla f(x+tp)^T p \, dt$$
$$\Rightarrow \nabla f(x+p) - \nabla f(x) = \int_0^1 \nabla^2 f(x+tp) \, p \, dt$$
$$\Rightarrow \nabla f(x_{k+1}) - \nabla f(x_k) = \int_0^1 \nabla^2 f(x_{k+1}) \, p_k \, dt$$
$$\Rightarrow \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k) p_k = \int_0^1 [\nabla^2 f(x_{k+1}) - \nabla^2 f(x_k)] p_k \, dt$$
$$\leq \frac{L}{2} \| p_k \|^2$$
$$= \frac{L}{2} \| x_{k+1} - x_k \|^2.$$

So if $x_{k+1} \approx x^*$ and $x_k \approx x^*$ (near x^*), then

$$\nabla f(x_{k+1}) \cong \nabla f(x_k) + \nabla^2 f(x_k) p_k$$

or, in other words,

$$\underbrace{\nabla f(x_{k+1}) - \nabla f(x_k)}_{\stackrel{def}{=}\, y_k} \cong \underbrace{\nabla^2 f(x_k)}_{\stackrel{def}{=}\, B_{k+1}} \underbrace{(x_{k+1} - x_k)}_{\stackrel{def}{=}\, s_k}.$$

Therefore, we can replace Newton Equation

$$P_k = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k) \longrightarrow \boxed{P_k = -\left[B_{k+1}\right]^{-1} \nabla f(x_k)}$$

Update Rules for B_{k+1}:

(Because we want to find cheap ways to compute B_{k+1})

(i) Symmetric Rank-1 Update

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{(y_k - B_k s_k)^T s_k}$$

(ii) BFGS:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{y_k^T s_k}.$$

Why these Updates?

- Maintain Symmetry
- Prefer low-rank update
- satisfies $B_{k+1} s_k = y_k$.

Barzilai - Borwein Method

- Steepest descent with a special step size
- Motivated from Newton but does not require Hessian.

Recall Newton method:

$$\nabla^2 f(x_k)\, p_k = -\nabla f(x_k)$$

$$\Longrightarrow \quad p_k = -\left[\nabla^2 f(x_k)\right]^{-1} \nabla f(x_k)$$

Idea: Replace $-\left[\nabla^2 f(x_k)\right]^{-1}$ by $-\alpha_k I$.

So we want

$$\nabla^2 f(x_k) \longleftarrow \alpha_k^{-1} I.$$

Let's also recall Quasi-Newton Method:

$$B_{k+1} s_k = y_k \qquad \begin{cases} s_k = x_{k+1} - x_k \\ y_k = \nabla f(x_{k+1}) - \nabla f(x_k) \end{cases}$$

So, we can choose α_k s.t.

$$\boxed{-(\alpha_k^{-1} I)\, s_{k-1} = y_{k-1}.}$$

How to choose α_k?

(i) $$\alpha_k^{-1} = \underset{\beta}{\operatorname{argmin}} \| s_{k-1}\beta - y_{k-1} \|^2$$

$$\Longrightarrow \alpha_k = \cancel{s_k}\ \frac{s_{k-1}^T s_{k-1}}{s_{k-1}^T y_{k-1}}$$

(ii) $$\alpha_k = \underset{\alpha}{\operatorname{argmin}} \| s_{k-1} - y_{k-1}\alpha \|^2$$

$$\Longrightarrow \alpha_k = \frac{s_{k-1}^T y_{k-1}}{y_{k-1}^T y_{k-1}}.$$

Gradient Projection Method

Recall steepest descent:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

<u>Claim:</u>

$$x_{k+1} = \underset{x}{\text{argmin}} \left\{ f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}$$

<u>Proof</u>: Take derivative:

$$\frac{\partial}{\partial x}(\cdot) = \nabla f(x_k) + \frac{1}{\alpha_k}(x - x_k) = 0$$

$$\Rightarrow \quad x = x_k - \alpha_k \nabla f(x_k).$$

<u>Implication:</u>

Gradient descent is a minimizer of a local linear model added with a quadratic penalty. (or think as we set $\nabla^2 f(x_k) \cong \frac{1}{\alpha_k} I$)

<u>Gradient Projection:</u>

if we assume that $x \in \Omega$, say $l \le x \le u$, then we can modify the min. as

$$x_{k+1} = \underset{x \in \Omega}{\text{argmin}} \left\{ f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2\alpha_k} \|x - x_k\|^2 \right\}.$$

$$\Rightarrow \boxed{x_{k+1} = P_\Omega\left(x_k - \alpha_k \nabla f(x_k) \right)}$$

where P_Ω is a projection operator that projects its input to Ω.

Example: GPSR algorithm (Figueiredo, Nowak, Wright)

GPSR tries to solve this problem

$$\min_{x} \frac{1}{2}\|y - Ax\|^2 + \lambda\|x\|_1$$

Its idea is to turn the problem into

$$\min_{u,v} \frac{1}{2}\| y - A(u-v)\|^2 + \lambda\left(\mathbf{1}^T u + \mathbf{1}^T v\right)$$
$$\text{s.t. } u \geq 0,\ v \geq 0.$$

$$\Longleftrightarrow \boxed{\begin{aligned} &\min_{z}\ f(z) = \frac{1}{2} z^T B z + c^T z \\ &\text{s.t. } z \geq 0. \end{aligned}}$$

where $z = \begin{bmatrix} u \\ v \end{bmatrix}$, $b = A^T y$, $c = \lambda \mathbf{1} + \begin{bmatrix} -b \\ b \end{bmatrix}$,

$$B = \begin{bmatrix} A^TA, & -A^TA \\ -A^TA, & A^TA \end{bmatrix}.$$

- The gradient projection is the following:

$$z_{k+1} = \left[z_k - \alpha_k \nabla f(z_k) \right]_+ ,$$

where $\nabla f(z_k) = B z_k + c$,

and $[\cdot]_+$ returns the positive entries only.

e.g. $[x_i]_+ = \begin{cases} x_i, & \text{if } x_i \geq 0 \\ 0, & \text{if } x_i < 0. \end{cases}$

The step size α_k is found using the line search.

- Variation: (Barzilai–Borwein)

$$z_{k+1} = z_k - \lambda_k p_k,$$

λ_k found by line search

where $p_k = z_k - z_{k-1}$

$$= \left[z_{k-1} - \alpha_k \nabla f(z_{k-1}) \right]_+ - z_{k-1}.$$

$\alpha_k = \frac{s_k^T s_k}{s_k^T B_{k+1} s_k}$ (BB step)

Proximal Gradient Methods

The proximal mapping of a <u>convex</u> function h is defined as

$$\text{prox}_h(x) = \underset{v}{\text{argmin}} \left\{ h(v) + \frac{1}{2}\|v - x\|^2 \right\}$$

For $\lambda h(x)$, it is easy to show that

$$\text{prox}_{\lambda h}(x) = \underset{v}{\text{argmin}} \left\{ h(v) + \frac{1}{2\lambda}\|v - x\|^2 \right\}.$$

<u>Examples</u>

(i) $h(x) = 0$. Then $\text{prox}_h(x) = x$

(ii) $h(x) = \begin{cases} 0, & \text{if } x \in \Omega \\ +\infty, & \text{if } x \notin \Omega, \end{cases}$

then

$$\text{prox}_h(x) = \underset{v \in \Omega}{\text{argmin}} \frac{1}{2}\|v - x\|^2 = P_\Omega(x).$$

if $\Omega = \{x \mid Ax = b\}$ and if A full rank, then,

$P_\Omega(x) = x - A^{\dagger}(Ax - b)$

Pseudo-inverse

(iii) $h(x) = \|x\|_1$. Then

$$\text{prox}_h(x) = \underset{v}{\text{argmin}} \left\{ \|x\|_1 + \frac{1}{2}\|v - x\|^2 \right\}$$

in general,

$$\text{prox}_{\lambda h}(x) = \underset{v}{\text{argmin}} \left\{ \|x\|_1 + \frac{1}{2\lambda}\|v - x\|^2 \right\} = S_\lambda(x) = \max\{|x| - \lambda, 0\}\,\text{sign}(x).$$

scalar x

(iv) $h(x) = -\log x$ (appears for Poisson regression)

$$\text{prox}_{\lambda h}(x) = \underset{v}{\text{argmin}}\left\{-\log v + \frac{1}{2\lambda}(v-x)^2\right\}$$

$$\frac{\partial}{\partial v}(\cdot) = \frac{-1}{v} + \frac{1}{\lambda}(v-x) = 0$$

$$\Rightarrow \quad -\lambda + v^2 - vx = 0$$

$$\Rightarrow \quad v = \frac{x + \sqrt{x^2+4\lambda}}{2}.$$

This technique applies to vector case if

$$h(x) = -\sum_{i=1}^{n} \log x_i .$$

(v) $h(x) = \begin{cases} 0, & x \in \Omega \\ +\infty, & x \notin \Omega. \end{cases}$

happens for probability-type of problems

$$\Omega = \{x \mid x \geq 0, \ x^T \mathbb{1} = 1\}$$

$$\text{prox}_{\lambda h}(x) = \underset{v}{\text{argmin}}\left\{\frac{1}{2}\|v - x\|^2\right\}$$
$$\text{s.t.} \quad x \geq 0, \quad x^T \mathbb{1} = 1.$$

$= (x - \lambda \mathbb{1})_+$, where λ is the solution of

$$\mathbb{1}^T(x - \lambda\mathbb{1}) = 1, \text{ ~~or~~}$$

i.e. $\displaystyle\sum_{i=1}^{n} \max\{x_i - \lambda, 0\} = 1.$

(This is a constrained problem. We will discuss it after we study the KKT condition & Lagrange multiplier.)

Consider an unconstrained optimization with two functions in the objective

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

- $f(x)$ is convex and differentiable (e.g. $\|Ax-y\|^2$)
- $h(x)$ is convex (e.g. $\lambda\|x\|_1$).

The proximal gradient method is

$$x_{k+1} = \text{prox}_{\alpha_k h}\left(x_k - \alpha_k \nabla f(x_k) \right)$$

step size (constant or line search)

What is it doing?

$$
\begin{aligned}
x_{k+1} &= \text{prox}_{\alpha_k h}\left(x_k - \alpha_k \nabla f(x_k) \right) \\
&= \operatorname*{argmin}_v \left\{ h(v) + \frac{1}{2\alpha_k} \| v - (x_k - \alpha_k \nabla f(x_k)) \|^2 \right\} \\
&= \operatorname*{argmin}_v \left\{ h(v) + \frac{1}{2\alpha_k} \left(\|v - x_k\|^2 + 2\alpha_k \nabla f(x_k)^T (v - x_k) + \cancel{\alpha_k^2 \|\nabla f(x_k)\|^2} \right) \right\} \\
&= \operatorname*{argmin}_v \left\{ h(v) + \cancel{2\alpha_k} \nabla f(x_k)^T (v - x_k) + \frac{1}{2\alpha_k} \|v - x_k\|^2 \right\} \\
&= \operatorname*{argmin}_v \left\{ h(v) + \underbrace{f(x_k) + \nabla f(x_k)^T (v - x_k) + \frac{1}{2\alpha_k} \|v - x_k\|^2}_{\text{local quadratic model around } x_k} \right\}
\end{aligned}
$$

Example:

(i) $h(x) = 0$. Then

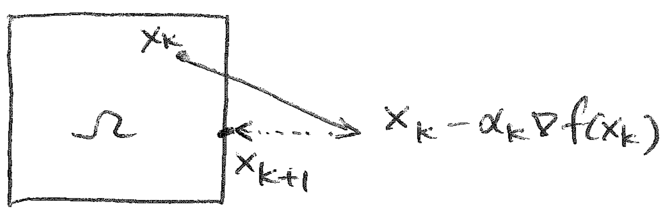
$$
\begin{aligned}
x_{k+1} &= \text{prox}_{\alpha_k h}\left(x_k - \alpha_k \nabla f(x_k)\right) \\
&= \operatorname*{argmin}_v \left\{ 0 - \frac{1}{2\alpha_k} \| v - (x_k - \alpha_k \nabla f(x_k)) \|^2 \right\} \\
&= x_k - \alpha_k \nabla f(x_k) \quad \text{See previous example (i)}
\end{aligned}
$$

So the proximal mapping is the classical <u>steepest descent</u>.

(ii) $h(x) = \begin{cases} 0, & x \in \Omega \\ +\infty, & x \notin \Omega. \end{cases}$ Then

$$
\begin{aligned}
x_{k+1} &= \text{prox}_{\alpha_k h}\left(x_k - \alpha_k \nabla f(x_k)\right) \\
&= P_{\Omega}\left(x_k - \alpha_k \nabla f(x_k)\right)
\end{aligned}
$$

So the proximal mapping is the <u>project gradient</u> method.



(iii) $h(x) = \|x\|_1$. Then

$$
\begin{aligned}
x_{k+1} &= \text{prox}_{\alpha_k h}\left(x_k - \alpha_k \nabla f(x_k)\right) \\
&= S_{\alpha_k}\left(x_k - \alpha_k \nabla f(x_k)\right) \\
&= \max\left\{ |x_k - \alpha_k \nabla f(x_k)| - \alpha_k, 0 \right\} \text{sign}\left\{ x_k - \alpha_k \nabla f(x_k) \right\}.
\end{aligned}
$$

Application: ISTA

$$\min_x f(x) + h(x), \text{ where}$$

$$\begin{cases} f(x) = \frac{1}{2}\|Ax - y\|^2 \\ h(x) = \lambda \|x\|_1 . \end{cases}$$

The proximal gradient method defines

$$x_{k+1} = \text{prox}_{\alpha_k h}\left(x_k - \alpha_k \nabla f(x_k)\right).$$

$$\nabla f(x_k) = A^T(Ax_k - y)$$

$$\begin{aligned} \text{prox}_{\alpha_k h}(x) &= \operatorname*{argmin}_v \left\{ \lambda \|x\|_1 + \frac{1}{2\alpha_k}\|v - x\|^2 \right\} \\ &= \operatorname*{argmin}_v \left\{ \|x\|_1 + \frac{1}{2\lambda\alpha_k}\|v - x\|^2 \right\} \\ &= S_{\lambda\alpha_k}(x). \end{aligned}$$

Convergence Rate of ISTA:

Assume f is differentiable & convex, and has Lipschitz continuous gradient:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

e.g. if $f(x) = \frac{1}{2}\|Ax - y\|^2$, then

$$\nabla f(x) = A^T(Ax - y)$$

$$\begin{aligned} \|\nabla f(x) - \nabla f(x')\| &= \|A^T(Ax - y) - A^T(Ax' - y)\| \\ &= \|A^TA(x - x')\| \\ &\leq \cancel{\|A^TA\|} \underbrace{\lambda_{max}(A^TA)}_{L} \|x - x'\| \end{aligned}$$

Theorem (Beck & Teboulle)

Let $\{x_k\}$ be a sequence generated by ISTA with either a constant step size or a line search, then for any $k \geq 1$, it holds that

$$F(x_k) - F(x^*) \leq \frac{\gamma L \|x_0 - x^*\|^2}{2k},$$

where $\gamma = 1$ for constant step size.
$\gamma = \eta_s$ for line search.

$F(x) \stackrel{\text{def}}{=} f(x) + h(x)$, is the overall objective function.

If we want

$$F(x_k) - F(x^*) < \epsilon,$$

then Theorem tells us that we need

$$\frac{\gamma L \|x_0 - x^*\|^2}{2k} < \epsilon$$

$$\Longrightarrow \quad k > \frac{\gamma L \|x_0 - x^*\|^2}{2\epsilon} \longleftarrow O\left(\frac{1}{\epsilon}\right).$$

Operations.