Sum of Square

For constant D, solved by interior method ✓.

$\min_{x\in D} f(x) \to \min_{y\in D'} g(y)$, $r(y)=x$

$f(x)=\sum_{\vec{\alpha}} C_{\vec\alpha}\prod_{i\in[d]} x_i^{\vec{\alpha}_i}$

$X_{\vec\alpha}=\prod_{i\in[d]} x_i^{\vec\alpha_i}$, $X=\{X_{\vec\alpha}\}_{\vec\alpha}$

$G(X)=\sum_{\vec\alpha} C_{\vec\alpha} X_{\vec\alpha}$, linear, convex

$h(x)=\sum_{\vec\alpha} C(h)_{\vec\alpha}\prod_{i\in[d]} x_i^{\vec\alpha_i}$

add constraint $\sum_{\vec\alpha} C(h^2)_{\vec\alpha} X_{\vec\alpha}\ge 0$

> polytope constraint

Theorem: $\boxed{\min_X G(X)}=\min_x f(x)$,

$x^*=(X^*_{(1,0,\dots,0)},\dots,X^*_{(0,0,\dots,1)})$

Pseudo-expectation: $X_{\vec\alpha}$ ☆

$\tilde{\mathbb{E}}_X g=\boxed{\sum_{\vec\alpha} C(g)_{\vec\alpha}X_{\vec\alpha}}$ ☆

Write the minimization: $\min \tilde{\mathbb{E}}_X g$,

for all polynomial h: $\tilde{\mathbb{E}}_X h^2\ge 0$

Infinite h? → only for $\deg(h^2)\le D$

$\|\vec\alpha\|_1>D$, $C(h^2)_{\vec\alpha}=0$

Let $M_D(X)$ be the matrix, that

$[M_D(X)]_{\vec\alpha,\vec\alpha'}=X_{\vec\alpha+\vec\alpha'}$, $\|\vec\alpha\|_1,\|\vec\alpha'\|_1\le\frac{D}{2}$

$\boxed{M_D(X)\succeq 0}$ $\sum_{\vec\alpha} C(h^2)_{\vec\alpha}X_{\vec\alpha}=c^T M_D(X)c$

for $c=(C(h)_{\vec\alpha})_{\vec\alpha}$, degree D SOS

Constrained: $\min f(x)$, s.t. $\{h_i(x)\ge 0\}_{i\in[m]}$.

still: $G(X)=\sum_{\vec\alpha} C_{\vec\alpha}X_{\vec\alpha}$

$\sum_{\vec\alpha} C(h_i h^2)_{\vec\alpha}X_{\vec\alpha}\ge 0$

Robust Linear Regression:

$\min_{a_i}\min_w \sum_{i\in[n]} a_i(\langle x^{(i)},w\rangle-y^{(i)})^2$

$a_i^2=a_i$, $\sum_i a_i\ge 1-\gamma$

Momentum

smoothness: $\|\nabla f(x)-\nabla f(y)\|\le L\|x-y\|$

Adjust the lr automatically

Weighted sum/average of past gradient

Instead of $x_{t+1}=x_t-\eta\nabla f(x_t)$

$x_{t+1}=x_t-\eta g_t$

Nesterov's: $z_{t+1}=x_t-\eta\nabla f(x_t)$

$x_{t+1}=(1-\gamma_t)z_{t+1}+\gamma_t z_t$

~~x_{t+1}~~ for small $\gamma>0$)

Heavy ball: $x_{t+1}=x_t-\eta g_t$, $g_t=\gamma\sum_{s\le t}(1-\gamma)^{t-s}\nabla f(x_s)$

we choose $\eta=\frac{1}{\gamma L}\gg\frac{1}{L}$

Gradient descent Lemma: $f(x_{t+1})\le f(x_t)-\frac{\eta}{2}\|\nabla f(x_t)\|_2^2$

Linear coupling: $s_{t+1}=x_t-\frac{1}{L}\nabla f(x_t)$

$l_{t+1}=l_t-\eta\nabla f(x_t)$, $\eta\gg\frac{1}{L}$

$x_{t+1}=(1-\tau)s_{t+1}+\tau l_{t+1}$

$\boxed{\text{In practice}}$ $x_{t+1}\approx x_t-\eta g_t$, $g_t=\gamma\sum_{s\le t}(1-\gamma)^{t-s}\nabla f(x_s)$

Proof: $T_{GD}\approx T^2_{AGD}$. $O(1/t^2)$

Second order

second order taylor:

$f(y)\le f(x)+\langle\nabla f(x),y-x\rangle+\frac{1}{2}(y-x)^T\nabla^2 f(x)(y-x)$

$\delta=-[\nabla^2 f(x)]^{-1}\nabla f(x)$

Newton's method: Lipschitz Hession

$x_{t+1}=x_t-\eta[\nabla^2 f(x_t)]^{-1}\nabla f(x_t)$ inefficient

Only local convergence: $\nabla f(x^*)=0,\nabla^2 f(x^*)\succ 0$

$\frac{\|x_{t+1}-x^*\|_2}{\sigma/2L}\le\left(\frac{\|x_t-x^*\|_2}{\sigma/2L}\right)^2$ quadratic convergence

Global convergence for "sandwich" function

$A\preceq\nabla^2 f(x)\preceq\alpha A$, $x_{t+1}=x_t-\eta[\nabla^2 f(x_0)]^{-1}\nabla f(x_t)$

$\|x_{t+1}-x^*\|_2\le(1-\eta/\alpha)\|x_t-x^*\|_2$ linear rate

$\frac{M}{\alpha}\preceq\nabla^2 f(x)\preceq\alpha M$, $x_{t+1}=x_t-\eta M^{-1}\nabla f(x_t)$

pre-condition matrix

Interior point method

Motivation: Constraint optimization

theorectically fast

$R(x)=\begin{cases}=+\infty, & x\in\partial D\\ \in(-\infty,\infty), & x\in D^\circ\end{cases}$

minimize $f(x)+\lambda R(x)$

Can not be smooth/Lipschitz. differentiable.

Lipschitzness of h: $\|\nabla h(x)\|_2$

smoothness of h: $\|\nabla^2 h(x)\|$ spectral.

Self-concordance with parameter ν

$\langle\nabla R(x),v\rangle^2\le\nu\, v^T\nabla^2R(x)v$

$|v^T\nabla^2R(x)v|^{3/2}\ge\frac{1}{2}\times\nabla^3R(x)(v,v,v)$

$\nabla^3 R(x)(v,v,v)=\frac{d^3}{dt^3}R(x+tv)|_{t=0}$

$\langle\nabla R(x),v\rangle=\frac{d}{dt}R(x+tv)|_{t=0}$

$v^T\nabla^2R(x)v=\frac{d^2}{dt^2}R(x+tv)$

change slowly curve.

are not "scaling invariant"

convex ✓, every convex D, ✓

R_1,R_2,ν selfconcordant

R_1+R_2, 2ν; $R_3=R_1(Ax+b)$, ν

Example: $D=\{x\in\mathbb{R}^d\mid\forall i\in[n],\ \langle a_i,x\rangle\le b_i\}$

$R(x)=-\sum_{i\in[n]}\log(b_i-\langle a_i,x\rangle)$

interior point method:

$\lambda_{t+1}=\lambda_t(1-\beta)$, $\beta\in(0,1)$

$x_{t+1}^*=\arg\min\{f(x)+\lambda_{t+1}R(x)\}$

by pre-conditioned gd with η.

Dikin's Ellipsoid:

$E_t=\{x\in\mathbb{R}^d\mid(x-x_t^*)^T\nabla^2R(x_t^*)(x-x_t^*)\le\frac{1}{8}\}$

$\frac{1}{4}\nabla^2R(x_t^*)\preceq\nabla^2F_{t+1}(x)/\lambda_{t+1}\preceq 4\nabla^2R(x_t^*)$

Self concordant implies sandwich in DE

Adagrad

$x_{t+1}=x_t-M^{-1}\nabla f(x_t)$

M: diagnal pre conditioning

coordinate of $\nabla f(x_t)$ large

scale it down

each coordinate has abs value one

gradient sign:

$x_{t+1}=x_t-\eta\,\mathrm{sign}(\nabla f(x_t))$

Adagrad: $x_{t+1}=x_t-\eta M^{-1}\nabla f(x_t)$

$M_t=\mathrm{diag}\left(\left\{\sqrt{\sum_{s\le t}\nabla f(x_s)_i^2}\right\}_{i=1}^d\right)$

more stable, compared to especially SGD

convergence ✓

no need to "η" "M"

Adagrad ⇔ mirror descent

$\phi_t(x)=\frac{1}{2}x^TM_tx$, $\nabla\phi_t(x)=M_tx$

$\nabla\phi_t(x_{t+1})=\nabla\phi_t(x_t)-\eta\nabla f(x_t)$

Adam: Adagrad + momentum.

~~No convergence~~ even in convex

$g_{t+1}=\gamma g_t+(1-\gamma)\nabla f(x_t)$

$s_{t+1}^2=\beta s_t^2+(1-\beta)[\nabla f(x_t)]^2$ ☆

$x_{t+1}=x_t-\eta\,\mathrm{diag}(s_{t+1})^{-1}g_{t+1}$

Distributed

1. $\min_w\frac{1}{|S_j|}\sum_{i\in S_j}l(h(x_i,w),y_i)+\boxed{R(w)}$

2. $\nabla f(w_t)=\sum_{j\in[m]}\nabla_j+\nabla R(w_t)$

per iter $O(md)$, $w\in\mathbb{R}^d$, gd convergence

ADMM: does not depend on smoothness / Lipschitzness

$\min_{\{w_j\}_{j\in[m]},w}\frac{1}{m}\sum_{j\in[m]}(f_j(w_j)+\lambda\|w_j-w\|_2^2)$ s.t. $w_j=w$

$\max_{\{d_j\}}\min_{\{w_j\},w}\frac{1}{m}\sum_{j\in[m]}(f_j(w_j)+\lambda\|w-w_j\|_2^2+\langle d_j,w_j-w\rangle)$

Locally: $w_j^{(t+1)}=\arg\min_{w_j}(f_j(w_j)+\lambda\|w_j-w\|_2^2+\langle d_j^{(t)},w_j-w\rangle)$

$w^{t+1}=\frac{\sum_j d_j^{(t)}}{2m\lambda}+\frac{1}{m}\sum w_j^{(t+1)}$

$d_j^{t+1}=d_j^t-\eta(w^{(t+1)}-w_j^{(t+1)})$ $O(md)$

Online optimization

env: pick $f_t: D\to\mathbb{R}$

act: chose $x_t\in D$. not knowing f_t.

env: $f_t(x_t)$, $\nabla f_t(x_t)$

env changing independent of act.

want act to adapt

$\sum_{t\in[T]}f_t(x_t)$ $\min_{x\in D}\sum_{t\in[T]}f_t(x)$

do as good as best fixed x

regret of player:

$R=\frac{1}{T}\sum_{t\in[T]}f_t(x_t)-\min_{x\in D}\sum f_t(x)$

if $f_t=f$, gradient descent, convex

if $\mathbb{E}[f_t(x)]=f(x)$ SGD

when f_t are convex: gd ✓

L-Lipschitz

$x^*=\arg\min\sum f_t(x)$, $\eta=\frac{\|x^*-x_0\|}{L\sqrt{T}}$

$R=\frac{1}{T}\sum_t f(x_t)-\frac{1}{T}\sum f_t(x^*)\le\frac{2L\|x^*-x_0\|_2}{\sqrt{T}}$

Projected: $x_{t+1}=\Pi_D(x_t-\eta\nabla f_t(x_t))$

Proof: three term MD

$f_t(x_t)\le f_t(x)+\frac{1}{2\eta}(\|x-x_t\|_2^2-\|x-x_{t+1}\|_2^2)$

$\frac{1}{T}\sum f_t(x_t)\le\frac{1}{T}\sum f_t(x)+\frac{1}{2\eta T}(\|x-x_0\|_2^2+\eta^2L^2)$

Stokes: for every $r\ge 0$, v randomly sampled unit vector

$\mathbb{E}_v\left[\frac{d}{r}f(x+rv)v\right]=\nabla\tilde{f}(x)$

$\approx\nabla f(x)$

$x_{t+1}=x_t-\frac{d}{r}f_t(x_t+rv)v$

scale with $\frac{d}{r}$, $V=\sum\|f_t-f_{t-1}\|$

Simulated Annealing

$f: \mathbb{R}^d \to \mathbb{R}$

$p(x) \propto e^{-\lambda f(x)}$

$\lambda \to 0$, uniform. temprature $\to \infty$

$\lambda \to +\infty$, global minimizer f

Metropolis-Hastings

$y_{t+1} \sim N(x_t, \sigma^2 I)$

$\alpha = \min\left\{\frac{p(y_{t+1})}{p(x_t)}, 1\right\}$

$x_{t+1} = \begin{cases} y_{t+1}, & \text{w.p. } \alpha \\ x_t, & \text{w.p. } 1-\alpha \end{cases}$

theorem: $t \to \infty$, $x_t \to \sim p(x)$

stationary distribution of Metropolis-Has

$\pi(x) = \int_{x'} p(x|x')\pi(x')\,dx'$

$= \int_{x'} \alpha_{x,x'}\, n(x|x')\pi(x')\,dx'$

$+ \pi(x)\int_z (1-\alpha_{z,x})\, n(z|x)\,dz$

$= \int_{x'} n(x|x')\alpha_{x',x}\pi(x)\,dx' + \dots$

$= \pi(x)\int_z n(z|x)\,dz = \pi(x)$.

Key idea: $Q \to P$. converge faster.

Simulate annealing.

First: sample $x_0 \sim D$. $\lambda_0 > 0$ small.

step: $\lambda_{t+1} = (1+\eta)\lambda_t$

M-H

$p_{\lambda_{t+1}}(x) \propto e^{-\lambda_{t+1} f(x)}$

Evolve the p: random $\to$ supports only on minimizer

no efficient rate guaratee. unless f convex

convergence ✓. step $\to \infty$. MH.

Evolutionary strategies.

$f: \mathbb{R}^d \to \mathbb{R}$

$\epsilon_1, \dots, \epsilon_N \sim N(0, I_{d\times d})$

$f_i = f(x_t + \sigma_t \epsilon_i)$

$x_{t+1} = x_t + \eta \frac{1}{N\sigma_t}\sum^N f_i \epsilon_i$

$f_i \uparrow$ higher weight for ϵ_i

do SGD on $F_t = f * g_t$

g_t: $N(0, \sigma_t^2 I)$.

$*$: $[f*g](x) = \int_y f(y)\,g(y-x)\,dy$

$\nabla F_t = f * \nabla g_t$ stoke's.

escape local minma when $\sigma_t \uparrow$

Non-convex

$f(x+\tau) = f(x) + \langle \nabla f(x), \tau\rangle + \frac{1}{2}\tau^T \nabla^2 f(x)\tau \pm O(\|\tau\|_2^3)$

$\nabla^2 f(x)$ PSD?

second order local minima: $\nabla f(x) = 0$, $\nabla^2 f(x) \succeq 0$

$f(x) = x^3$. $x=0$ local minima

Hessian descent. not in practice

: $y_{t+1} = x_t - \eta \nabla f(x_t)$

unit vector v, corresponding to eigen vector of $\nabla^2 f(x)$ with smallest eigen value.

$z_{t+1,1} = x_t - \eta v$, $z_{t+1,2} = x_t + \eta v$

$x_{t+1} = \arg\min\{f(y_{t+1}), f(z_{t+1,1}), f(z_{t+1,2})\}$

γ-Hessian Lipschitz:

$f(x+\tau) \le f(x) + \langle \nabla f(x), \tau\rangle + \frac{1}{2}\tau^T \nabla^2 f(x)\tau + \gamma\|\tau\|_2^3$

$\frac{1}{2}(f(z_{t+1,1}) + f(z_{t+1,2})) \le f(x_t) - \frac{\eta^2}{4}\delta$

Hessian negative. locally high non-convex

$f(y_{t+1}) \le f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|_2^2$

Noisy GD ✓ efficiently find second order minima

$x_{t+1} = x_t - \eta \nabla f(x_t)$

$x_{t+1} \leftarrow x_{t+1} + \xi_{t+1} \sim N(0, \sigma^2 I)$

$\nabla f(x_{t+1}) = \nabla f(x_t) - \eta \nabla^2 f(x_t)\nabla f(x_t) + O(\eta^2)$

$\approx (I - \eta \nabla^2 f(x_t))\nabla f(x_t)$

when $\nabla^2 f(x_t)$ not PSD.

$\|I - \eta \nabla^2 f(x_t)\|_2 > 1$, gradient large.

Only stop when. gradient small, Hessian PSD

$\|\nabla f(x_t)\|_2 \le \rho$. $|\langle v, \nabla f(x_{t+1})\rangle|$ increasing

geometry rate

~~convergence rate~~ poly$(1/\epsilon, 1/\delta)$

dimension free. gradient $< \epsilon$, Hessian $\succeq \delta I$

BO

"local minima can not efficiently

spirit of BO: Graduate Student Descent

No need of computing gradient, global optimization.

low dimension ✓ high inefficient.

analog of Ellipsoid algorithm.

Generic routine:

$f: \mathbb{R}^d \to \mathbb{R}$, convex or not

step t: compute a distribution P_t over $f: \mathbb{R}^d \to \mathbb{R}$. using $f(x_0), \dots f(x_t)$

find x_{t+1}, maximize acquisition func associated with P_t.

GP: given kernel $k(x,y)$: $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

$k(x,y) = \langle x, y\rangle$, $k(x,y) = e^{-\frac{\|x-y\|_2^2}{2\sigma^2}}$

measures inverse distance.

Caussian kernel, closer, larger

Define $V_t(x) = (k(x, x_0), \dots k(x, x_t))^T$

$F_t = (f(x_0), f(x_1) \dots f(x_t))^T$

$[M_t]_{i+1, j+1} = k(x_i, x_j)$

$P_t(x) \sim N(\mu_t(x), \sigma_t^2(x))$ ☆

where $\mu_t(x) = V_t(x)^T M_t^{-1} F_t$

$\sigma_t^2(x) = k(x,x) - V_t(x)^T M_t^{-1} V_t(x)$

$M_t(x) = \begin{pmatrix} k(x,x) & V_t(x)^T \\ V_t(x) & M_t \end{pmatrix} \to$ PSD

$\mu_t(x_j) = f(x_j)$, $\sigma_t^2(x_j) = 0$

$x_{t+1} = \arg\max_x \mathbb{E}_{g\sim p_t}[[g(x) - f_t^*]^+]$

$EI_t(x) = \mathbb{E}_{g\sim p_t}[[g(x) - f_t^*]^+]$

acqusition function

$[z]^+ = \mathrm{ReLU}(z)$

Using (noisy) gradient descent

compute maximizer

cheap for $EI_t(x)$

optimistic exploration

theory ~~unknown~~.

Over-parameterization

matrix sensing

$\min_{U,V} \frac{1}{N}\sum_i (\langle A_i, UV^T\rangle - y_i)^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2$

$\min_M \frac{1}{N}\sum(\langle A_i, M\rangle - y_i)^2 + \lambda\|M\|_*$

non convex matrix product equivalent to nuclear norm

$\min_{w\in\mathbb{R}^d} f(w) \to \min_{w\in\mathbb{R}^D} f(w)$, $D \gg d$

theorem. all the second order minia U, V has objective $\le 1.1\epsilon$, when $R \gg r$

$U, V \in \mathbb{R}^{d\times R}$

$U^*, V^* \in \mathbb{R}^{d\times r}$, ξ

$R \ge \mathrm{poly}(1/\epsilon)\, r$, $O(\mathbb{E}[\|\xi\|_F^2]) \le 0.01\epsilon$

proof: $f(U,V) \ge 1.1\epsilon$, $f(U^*, V^*) < \epsilon$

$\mathbb{E}_{U' \sim p_U, V' \sim p_V} f((1-\eta^2)U + \sqrt{2\eta^2 - \eta^4}\,U', (1-\eta^2)V + \sqrt{2\eta^2 - \eta^4}\,V') \le f(U,V) - 0.05\eta^2\epsilon$

contradict with $\nabla f(U,V) = 0$, $\nabla^2 f(U,V) \succeq 0$ $\to f(U,V) - O(\eta^3)$

Sampling, generative model.

generate from same distribut

diffusion approach: $q_0 \to q^*$

Brownian motion. $X_0 \sim q^*$

$dX_t = -X_t\,dt + \sqrt{2}\,dB_t \to$ Gaussian noise

As $t \to \infty$, $X_t \to N(0,1)$

$X_t = e^{-t}X_0 + \sqrt{1-e^{-2t}}\,y$

Forward: $q_0 = q^*$, $q_T \to$ Gaussian

$p_t(x) = q_{T-t}(x)$ sufficient large T

satisfies, $\frac{\partial q_t(x)}{\partial t} = \langle \nabla, x q_t(x)\rangle + \langle \nabla^2 q_t(x), E\rangle$

Backward satisfies

$\frac{\partial p_t(x)}{\partial t} = \frac{\partial q_{T-t}(x)}{\partial t} = -\langle \nabla, x q_{T-t}(x)\rangle - \langle \nabla^2 q_{T-t}(x), E\rangle$

$dY_t = Y_t + 2\nabla \ln p_t(Y_t)\,dt + \sqrt{2}\,dB_t$ $(Y_t \sim p_t)$

due to $2\nabla \ln p_t(Y) = \frac{2\nabla p_t(Y)}{p_t(Y)}$

only $x \sim q^* = q_0$, score estimation to compute $\nabla \ln q_t(X_t)$:

minimize: $\min_s \mathbb{E}_{X_t}\|s(X_t) - \nabla \ln q_t(X_t)\|_2^2$

Here. $X_0 \sim q_0 = q^*$, $X_t = e^{-t}X_0 + \sqrt{1-e^{-2t}}\,y$

Stein's formula $\int \nabla \cdot v\,dy = \int \langle y, v(y)\rangle\,dy$, $y \sim N(0, \sigma^2)$

DDPM:

$\min_s \mathbb{E}_{X_t}\|s_t(X_t) + \frac{1}{\sqrt{1-e^{-2t}}}y\|_2^2$

where $X_0 \sim q^*$ target distribution

$X_t = e^{-t}X_0 + \sqrt{1-e^{-2t}}\,y$

then use backward diffusion, starting $Y_0 \sim N(0,1)$

$dY_t = Y_t + 2s_{T-t}(Y_t)\,dt + \sqrt{2}\,dB_t$

E